STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

23 oktober 2020 9–16

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

Uppgift 1

a) Minsta kvadrat-skattningarna av $\tilde{\alpha}$ och β ges av

$$\begin{array}{rcl} \hat{\tilde{\alpha}} & = & \sum_i Y_i/25 = 48.0/25 = 1.92, \\ \hat{\beta} & = & \sum_i Y_i(x_i - \bar{x})/\sum_i (x_i - \bar{x})^2 = 83.0/2800 = 0.0296. \end{array} \tag{1}$$

b) Det förväntade priset på den lägenhet Lisa vill köpa är

$$\mu = E(Y) = \tilde{\alpha} + (90 - \bar{x})\beta = \tilde{\alpha} + 10\beta,$$

där vi i sista ledet utnyttjade $\bar{x} = \sum_i x_i/25 = 2000/25 = 80$. Detta väntevärde skattas med

$$\hat{\mu} = \hat{\tilde{\alpha}} + 10\hat{\beta}. \tag{2}$$

Vidare är Y oberoende av $\hat{\mu}$, eftersom Y inte ingår i skattningen av modellens parametrar. Härur följer att

$$\begin{array}{rcl} \mathrm{Var}(Y - \hat{\mu}) & = & \mathrm{Var}(Y) + \mathrm{Var}(\hat{\mu}) \\ & = & \sigma^2 + \mathrm{Var}(\hat{\alpha}) + 10^2\mathrm{Var}(\hat{\beta}) \\ & = & \sigma^2 + \sigma^2/25 + 10^2\sigma^2/\sum_i (x_i - \bar{x})^2 \\ & = & \sigma^2(1 + 1/25 + 100/2800) \\ & = & 1.0757\sigma^2. \end{array}$$

c) Antalet frihetsgrader för att skatta feltermernas varians är 25-2=23. Av detta följer att

$$\hat{\sigma} = \sqrt{\mathrm{Kvs(Residual)}/23} = \sqrt{0.30/23} = 0.1142. \tag{3}$$

Vidare är

$$\hat{\mu} = \hat{\tilde{\alpha}} + 10\hat{\beta} = 1.92 + 10 \cdot 0.0292 = 2.216.$$

Eftersom $23\hat{\sigma}^2/\sigma^2 \sim \chi^2(23)$ är oberoende av prediktionsfelet $Y - \hat{\mu} \sim N(0, 1.0757\sigma^2)$ så följer att

$$\frac{(Y-\hat{\mu})/\sqrt{1.0757\sigma^2}}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{Y-\hat{\mu}}{\sqrt{1.0757}\cdot\hat{\sigma}} \sim t(23)$$

har en t-fördelning med 23 frihetsgrader. Ett 95% prediktionsintervall för Y är därför

$$\begin{aligned} I_Y &= (\hat{\mu} - t_{0.025}(23)\sqrt{1.0757}\hat{\sigma}, \hat{\mu} + t_{0.025}(23)\sqrt{1.0757}\hat{\sigma}) \\ &= (2.216 - 2.0687 \cdot \sqrt{1.0757} \cdot 0.1142, 2.216 + 2.0687 \cdot \sqrt{1.0757} \cdot 0.1142) \\ &= (1,971, 2.462) \end{aligned}$$

Med andra ord kommer lägenheten med 95% sannolikhet att säljas för mellan 1.97 och 2.46 Mkr.

Uppgift 2

a) Låt $\boldsymbol{Y} = (Y_1, \ldots, Y_{25})^T$ vara observationsvektorn, $\mathbf{1} = (1, \ldots, 1)^T$ en kolumnvektor av längd 25. Vidare låter vi

$$\boldsymbol{X} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 \\ \vdots & \vdots \\ x_{1,25} - \bar{x}_1 & x_{2,25} - \bar{x}_2 \end{pmatrix} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$$

beteckna x-delen av designmatrisen, med centrerade kolumner $\boldsymbol{x}_1$ och $\boldsymbol{x}_2$. På matrisform skrivs den multipla linjära regressionsmodellen som

$$\boldsymbol{Y} = \tilde{\alpha}\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

där $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ innehåller de två effektparametrarna och $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{25})^T$ är feltermsvektorn.

b) Låt $\mathrm{Var}(\hat{\beta}_1)$ och $\mathrm{Var}_0(\hat{\beta}_1)$ ange variansen för skattningen av β_1 under den multipla linjära regressionsmodellen, respektive den enkla linjära regressionsmodellen från uppgift 1. Låt vidare s_{ij} beteckna elementet i $\boldsymbol{S}$ från rad i och kolumn j. Varationsinflationsfaktorn vid skattning av β_1 ges av

$$\begin{aligned} \mathrm{VIF} &= \mathrm{Var}(\hat{\beta}_1)/\mathrm{Var}_0(\hat{\beta}_1) \\ &= \sigma^2(\boldsymbol{S}^{-1})_{11}/(\sigma^2 s_{11}^{-1}) \\ &= \boldsymbol{S}_{11}^{-1} s_{11} \\ &= s_{11}s_{22}/(s_{11}s_{22} - s_{12}^2) \\ &= 2800 \cdot 6.0/(2800 \cdot 6.0 - 60^2) \\ &= 1.273, \end{aligned}$$

där vi i fjärde ledet utnyttjade formeln för att invertera en 2×2-matris.

c) Låt

$$\begin{aligned}\hat{\mu}_i &= \hat{\bar{\alpha}} + \hat{\beta}_1(x_{1i} - \bar{x}_1) + \hat{\beta}_2(x_{2i} - \bar{x}_2) \\ &= \bar{Y} + \hat{\beta}_1(x_{1i} - \bar{x}_1) + \hat{\beta}_2(x_{2i} - \bar{x}_2)\end{aligned}$$

vara skattningen av $\mu_i = E(Y_i)$ för den multipla linjär regressionsmodellen, där vi i sista ledet utnyttjade att $\hat{\bar{\alpha}} = \bar{Y}$, som enligt uppgift 1 har värdet 1.92. Förklaringsgraden ges av

$$R^2 = \frac{\text{Kvs(Regression)}}{\text{Kvs(Total)}} = \frac{\sum_i(\hat{\mu}_i - \bar{Y})^2}{\sum_i(Y_i - \bar{Y})^2}. \tag{4}$$

Låt $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_{25})^T$ vara vektorn av skattade väntevärden för responsvariablerna. Vi skriver om täljaren i (4) som

$$\begin{aligned}\text{Kvs(Regression)} &= \|\hat{\boldsymbol{\mu}} - \bar{Y}\mathbf{1}\|^2 \\ &= \|\hat{\beta}_1\boldsymbol{x}_1 + \hat{\beta}_2\boldsymbol{x}_2\|^2 \\ &= \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 \\ &= \hat{\boldsymbol{\beta}}^T\boldsymbol{S}\hat{\boldsymbol{\beta}} \\ &= \hat{\beta}_1^2 s_{11} + 2\hat{\beta}_1\hat{\beta}_2 s_{12} + \hat{\beta}_2^2 s_{22} \\ &= 0.02^2 \cdot 2800 + 2 \cdot 0.02 \cdot 0.2 \cdot 60 + 0.2^2 \cdot 6.0 \\ &= 1.84,\end{aligned}$$

där vi i tredje ledet införde $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^T$ och i fjärde ledet utnyttjade att $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}$. För nämnaren i (4) utnyttjar vi uppgift 1 och skriver

$$\begin{aligned}\text{Kvs(Total)} &= \text{Kvs(Residual)}_{\text{Uppg 1}} + \text{Kvs(Regression)}_{\text{Uppg 1}} \\ &= 0.30 + \hat{\beta}^2 \textstyle\sum_i (x_i - \bar{x})^2 \\ &= 0.30 + 0.0296^2 \cdot 2800 \\ &= 2.7532.\end{aligned}$$

Genom att bilda kvoten av de två sista uttrycken får vi slutligen

$$R^2 = \frac{1.84}{2.7532} = 0.668.$$

Uppgift 3

a) Låt Y_{ijkl} beteckna BMI för person $l \in \{1, 2\}$ med nivån $i \in \{1, 2\}$ på kostfaktorn, nivå $j \in \{1, 2, 3\}$ på livsstilsfaktorn och nivå $k \in \{1, 2, 3\}$ på den genetiska faktorn. I en typ I variansanalysmodell skrivs BMI för denna person som

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \varepsilon_{ijkl},$$

där μ är det genomsnittliga väntevärdet för alla personer i undersökningen, α_i är huvudeffekten för kost, β_j är huvudeffekten för livsstil, γ_k den genetiska huvudeffekten, $(\alpha\beta)_{ij}$ samspelet mellan kost och livsstil samt ε_{ijkl} en felterm. Det antas att feltermerna är oberoende och normalfördelade med väntevärde

0 och varians σ^2. För att undvika överparametrisering så införs restriktioner $\alpha_1 + \alpha_2 = 0$, $\beta_1 + \beta_2 + \beta_3 = 0$, $\gamma_1 + \gamma_2 + \gamma_3 = 0$ samt 4 linjärt oberoende restritioner $\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$ (totalt 5 summor men endast 4 av dem är linjärt oberoende).

b) Vi fyller ut variansanalystabellen genom att ange antal frihetsgrader och medelkvadratsumman för varje variationskälla:

Variationskälla	Kvs	f	Mkvs = Kvs/f
Kost	6.0	2-1=1	6.0
Livsstil	12.0	3-1=2	6.0
Genetisk	5.0	3-1=2	2.5
Samspel Kost och Livsstil	10.0	(2-1)(3-1)=2	5.0
Residual	28.0	28	1.0
Totalt	61.0	$N - 1 = 35$	

För att testa samspelet mellan kost- och livsstilsfaktorerna bildar vi

$$\text{F-kvot} = \frac{\text{Mkvs(Samspel)}}{\text{Mkvs(Residual)}} = \frac{5.0}{1.0} = 5.0 > F_{0.05}(2, 28) = 3.340.$$

Nollhypotesen att det saknas samspel mellan kost och livsstil, vad gäller inverkan på BMI, förkastas alltså på signifikansnivån 5%.

c) För att testa om den genetiska faktorn har någon signifikant inverkan på BMI så bildar vi

$$\text{F-kvot} = \frac{\text{Mkvs(Genetisk)}}{\text{Mkvs(Residual)}} = \frac{2.5}{1.0} = 2.5 < F_{0.05}(2, 28) = 3.340.$$

Nollhypotesen att den genetiska faktorn saknar inverkan på BMI förkastas alltså inte på nivån 5%.

Uppgift 4

a) Vi inför observationsvektorn $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^T = (5.1, -5.2, 1.2, -0.9)^T$ för de fyra mätningarna med balansvågen. De kan beskrivas med hjälp av en allmän linjär modell

$$\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}.$$

b) Minsta kvadrat-skattningen av modellparametrarna ges av

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \frac{1}{4}\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} 3.1 \\ 2.05 \end{pmatrix},$$

där vi i tredje ledet utnyttjade att $\boldsymbol{A}^T\boldsymbol{A} = 4\boldsymbol{I}_2$, där $\boldsymbol{I}_2$ är enhetsmatrisen av ordning 2. Kovariansmatrisen för skattningen av parametervektorn ges av

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \frac{\sigma^2}{4}\boldsymbol{I}_2 = \frac{\sigma^2}{4}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

Det innebär att $\hat{\theta}_1$ och $\hat{\theta}_2$ är oberoende med $\mathrm{Var}(\hat{\theta}_1) = \mathrm{Var}(\hat{\theta}_2) = \sigma^2/4$.

c) Vi kan utan inskräkning anta att Sten börjar med n mätningar med objekt 1 i skål A, medan skål B hålls tom. Då fås en skattning $\tilde{\theta}_1$ av θ_1 som är medelvärdet av dessa n mätningar, med $\mathrm{Var}(\tilde{\theta}_1) = \sigma^2/n$. Därefter utför Sten m mätningar med objekt 2 i skål A, medan skål B hålls tom. Medelvärdet av dessa mätningar ger en skattning $\tilde{\theta}_2$ av θ_2 med $\mathrm{Var}(\tilde{\theta}_2) = \sigma^2/m$. Totalt utför alltså Sten $n+m$ mätningar. För få samma varians som i Eriks försök så krävs $n = m = 4$, det vill säga totalt 8 mätningar.

d) Vi börjar med att skatta feltermernas varians σ^2. Eftersom kvadratsumman för residualerna är given så följer att

$$\hat{\sigma}^2 = \frac{\mathrm{Kvs(Residual)}}{N-k} = \frac{0.05}{4-2} = 0.025,$$

där vi i andra steget utnyttjade att antalet modellparametrar är $k = 2$ och antalet observationer $N = 4$. Konfidensregionen för $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ ges av

$$\begin{aligned} E &= \{\boldsymbol{\theta} = (\theta_1,\theta_2)^T; (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T(\boldsymbol{A}^T\boldsymbol{A})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})/(k\hat{\sigma}^2) \le F_{0.05}(k, N-k)\} \\ &= \{\boldsymbol{\theta}; \|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|^2 \le F_{0.05}(2,2)\hat{\sigma}^2/2\} \\ &= \{\boldsymbol{\theta}; (\theta_1-3.1)^2 + (\theta_2-2.05)^2 \le 19.0\cdot 0.025/2\} \\ &= \{\boldsymbol{\theta}; (\theta_1-3.1)^2 + (\theta_2-2.05)^2 \le 0.2375\}, \end{aligned}$$

där vi i andra ledet utnyttjade att $\boldsymbol{A}^T\boldsymbol{A} = 4\boldsymbol{I}_2$.

Uppgift 5

a) Vi kan skriva försöket som en allmän linjär modell

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

med parametervektor $\boldsymbol{\theta} = (\mu, \bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{E})^T$, designmatris

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

responsvektor $\boldsymbol{Y} = (11, 16, 18, 14, 18, 12, 21, 26)^T$ och feltermsvektor $\boldsymbol{\varepsilon} = (\varepsilon_{----+}, \ldots, \varepsilon_{+++++})^T$. Minsta kvadrat-skattningen av parametervektorn ges av

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \frac{1}{8}\boldsymbol{A}^T\boldsymbol{Y} = \begin{pmatrix} 17.0 \\ 2.5 \\ 0.0 \\ 2.75 \\ 2.25 \\ 1.5 \end{pmatrix} = \begin{pmatrix} \hat{\mu} \\ \hat{A} \\ \hat{B} \\ \hat{C} \\ \hat{D} \\ \hat{E} \end{pmatrix}.$$

b) Vi söker ett konfidensintervall för linjärkombinationen $\xi = \bar{A} + \bar{B} + \bar{C} + \bar{D} + \bar{E} = \boldsymbol{c}^T\boldsymbol{\theta}$ av parametrarna, med $\boldsymbol{c} = (0, 1, 1, 1, 1, 1)^T$. Från deluppgift a) får skattningen

$$\hat{\xi} = \boldsymbol{c}^T\hat{\boldsymbol{\theta}} = \hat{A} + \hat{B} + \hat{C} + \hat{D} + \hat{E} = 2.5 + 0.0 + 2.75 + 2.25 + 1.5 = 9.0$$

av ξ. Vidare gäller att

$$\mathrm{Var}(\hat{\xi}) = \sigma^2\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{c} = \frac{\sigma^2}{8}\boldsymbol{c}^T\boldsymbol{c} = \frac{5\sigma^2}{8}.$$

Enligt uppgift är $\hat{\sigma}^2 = \mathrm{Mkvs(Residual)} = 0.5$. Eftersom antalet frihetsgrader för residualerna är $N - k = 8 - 6 = 2$, där $N = 8$ är antalet observationer och $k = 6$ antalet parametrar, så följer att ett 95% konfidensintervall för ξ ges av

$$\begin{aligned} &(\hat{\xi} - \sqrt{\tfrac{5\hat{\sigma}^2}{8}}t_{0.025}(2), \hat{\xi} + \sqrt{\tfrac{5\hat{\sigma}^2}{8}}t_{0.025}(2)) \\ &= (9.0 - \sqrt{\tfrac{5}{16}} \cdot 4.3027, 9.0 + \sqrt{\tfrac{5}{16}} \cdot 4.3027) \\ &= (6.595, 11.405). \end{aligned}$$

c) Eftersom antalet faktorer är 5, så har vi $2^5 = 32$ effekter i den fullständiga modellen, enheten I, 5 huvudeffekter, 10 samspel av ordning 2, 10 samspel av ordning 3, 5 samspel av ordning 4 samt ett samspel av ordning 5. För ett 2^{5-2}-försök innebär kopplingsmönstret att dessa effekter delas in i 8 grupper med fyra effekter i varje grupp, där effekterna inom varje grupp inte kan särskiljas. Prövning visar att fjärde ordningens samspel, dvs $ABCD$, är kopplat till enheten, eftersom elementvis produkt av motsvarande fyra kolumner i designmatrisen ger

$$ABCD = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = I.$$

Likaså är ABE kopplat till enheten, eftersom

$$ABE = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = I.$$

Av detta följer att $I = I \cdot I = (ABCD)(ABE) = A^2B^2CDE = CDE$ också är kopplat till enheten. Således har vi kopplingsmönstret

$$I = ABCD = ABE = CDE$$

för den grupp där enheten ingår. I resterande sju grupper återfinner vi huvudeffekter och andra ordningens samspel enligt

$$\begin{aligned} AB &= AB(ABCD) = CD = AB(ABE) = E, \\ AC &= AC(ABCD) = BD, \\ AD &= AD(ABCD) = BC, \\ AE &= AE(ABE) = B, \\ BE &= BE(ABE) = A, \\ CE &= CE(CDE) = D, \\ DE &= DE(CDE) = C, \end{aligned}$$

Endast två av dessa grupper saknar en huvudeffekt, och i dessa två grupper ingår andra ordningens samspel $AC = BD$ respektive $AD = BC$. Således kan vi endast ta med två samspel av ordning 2 i modellen; $\{AC, AD\}$, $\{AC, BC\}$, $\{BD, AD\}$ eller $\{BD, BC\}$. Dock: Om vi tar med två samspel av ordning 2 får vi $k = N = 8$ parametrar i modellen, så att feltermsvariansen inte kan skattas. Om man även vill kunna skatta feltermsvariansen kan man därför högst ta med ett samspel av ordning två i modellen.