STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

23 oktober 2023 14–19

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

Uppgift 1

a) Vi börjar med att beräkna minsta kvadrat-skattningarna av de två regressionsparametrarna $\tilde{\alpha}$ och β. De ges av

$$\begin{aligned} \hat{\tilde{\alpha}} &= \textstyle\sum_1^{10} y_i/10 = 54.8/10 = 5.48, \\ \hat{\beta} &= \textstyle\sum_1^{10} y_i(x_i - \bar{x}) / \sum_1^{10}(x_i - \bar{x})^2 = 2.8/20 = 0.14, \end{aligned}$$

där vi för skattningen av lutningsparametern β utnyttjade att $\bar{x} = 3$ och

$$\begin{aligned} \textstyle\sum_1^{10}(x_i - \bar{x})^2 &= 2\left[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2\right] \\ &= 2\,(4 + 1 + 0 + 1 + 4) \\ &= 20. \end{aligned}$$

Det ger en skattning

$$\hat{\mu}(6) = \hat{\tilde{\alpha}} + (6 - \bar{x})\hat{\beta} = \hat{\tilde{\alpha}} + 3\hat{\beta} = 5.48 + 3 \cdot 0.14 = 5.90$$

av utandningsvolymen hos en patient som tar 6 mg av medicinen.

b) Från variansanalystabellen får vi först att

$$\text{Kvs(Residual)} = \text{Kvs(Total)} - \text{Kvs(Regression)} = 1.216 - 0.392 = 0.824.$$

Eftersom antalet frihetsgrader för Residual är 10-2=8, följer att

$$\hat{\sigma}^2 = \text{Mkvs(Residual)} = \frac{\text{Kvs(Residual)}}{8} = \frac{0.824}{8} = 0.103$$

är en väntevärdesriktig skattning av σ^2.

c) Prediktionsintervallet för utandningsvolymen Y hos en patient som tagit dosen 6 mg av medicinen, ges av

$$\begin{aligned} I_Y &= (\hat{\mu}(6) - t_{0.025}(8) \cdot d, \hat{\mu}(6) + t_{0.025}(8) \cdot d) \\ &= (5.90 - 2.306 \cdot d, 5.90 + 2.306 \cdot d), \end{aligned} \tag{1}$$

där medelfelet

$$\begin{aligned} d &= \hat{\sigma}\sqrt{1+\tfrac{1}{10}+\tfrac{(6-\bar{x})^2}{\sum_1^{10}(x_i-\bar{x})^2}} \\ &= \sqrt{0.103}\cdot\sqrt{1+\tfrac{1}{10}+\tfrac{9}{20}} \\ &= 0.3996 \end{aligned} \tag{2}$$

är en skattning av standardavvikelsen av prediktionsfelet, det vill säga en skattning av

$$\begin{aligned} \sqrt{\mathrm{Var}\left[Y-\hat{\tilde{\alpha}}-(6-\bar{x})\hat{\beta}\right]} &= \sqrt{\mathrm{Var}(Y)+\mathrm{Var}(\hat{\tilde{\alpha}})+(6-\bar{x})^2\mathrm{Var}(\hat{\beta})} \\ &= \sigma\cdot\sqrt{1+\tfrac{1}{10}+\tfrac{(6-\bar{x})^2}{\sum_1^{10}(x_i-\bar{x})^2}}. \end{aligned}$$

Insättning av (2) i (1) ger ett prediktionsintervall

$$I_Y = (5.90 - 2.306\cdot 0.3996, 5.90 + 2.306\cdot 0.3996) = (4.98, 6.82).$$

Uppgift 2

a) Nollhypotesen att endast livsstilsfaktorer påverkar ämnesomsättningen kan skrivas som

$$H_0 : \beta_4 = \beta_5 = 0.$$

Det svarar parametervektor $\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2, \beta_3, 0, 0)^T$ där de två sista komponenterna sätts till 0. Vi kan därför skriva hypotesmodellen på formen

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \boldsymbol{B}\boldsymbol{\lambda}.$$

b) Vi har att

$$\begin{aligned} R_0^2 &= \textstyle\sum_{i=1}^{30}(\hat{\mu}_i-\bar{Y})^2 / \sum_{i=1}^{30}(Y_i-\bar{Y})^2, \\ R_1^2 &= \textstyle\sum_{i=1}^{30}(\hat{\tilde{\mu}}_i-\bar{Y})^2 / \sum_{i=1}^{30}(Y_i-\bar{Y})^2. \end{aligned} \tag{3}$$

c) Det följer av (3) att

$$R_0^2 - R_1^2 = \frac{\sum_i(\hat{\mu}_i-\bar{Y})^2 - \sum_i(\hat{\tilde{\mu}}_i-\bar{Y})^2}{\sum_i(Y_i-\bar{Y})^2} = \frac{\sum_i(\hat{\mu}_i-\hat{\tilde{\mu}}_i)^2}{\sum_i(Y_i-\bar{Y})^2} \tag{4}$$

och

$$1 - R_0^2 = \frac{\sum_i(Y_i-\bar{Y})^2 - \sum_i(\hat{\mu}_i-\bar{Y})^2}{\sum_i(Y_i-\bar{Y})^2} = \frac{\sum_i(Y_i-\hat{\mu}_i)^2}{\sum_i(Y_i-\bar{Y})^2}. \tag{5}$$

Dessa två ekvationer inses lättast genom att införa observationsvektorn $\boldsymbol{Y} = (Y_1, \ldots, Y_{30})^T$, medelvärdesvektorn $\bar{Y} = (\bar{Y}, \ldots, \bar{Y})^T$, samt de skattade väntevärdesvektorerna $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_N)^T$ och $\hat{\hat{\boldsymbol{\mu}}} = (\hat{\hat{\mu}}_1, \ldots, \hat{\hat{\mu}}_N)^T$ under grund- och hypotesmodellen. I ekvation (4) utnyttjades att $\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}$ och $\hat{\hat{\boldsymbol{\mu}}} - \bar{Y}$ är ortogonala vektorer, samt i (5) att $\boldsymbol{Y} - \hat{\boldsymbol{\mu}}$ och $\hat{\boldsymbol{\mu}} - \bar{Y}$ är ortogonala.

Eftersom $N = 30$, grundmodellen har $k = 6$ parametrar och hypotesmodellen $l = 4$ parametrar, följer att

$$\begin{aligned} \text{F-kvot} &= \frac{\sum_i (\hat{\mu}_i - \hat{\hat{\mu}}_i)^2/(k-l)}{\sum_i (Y_i - \hat{\mu}_i)^2/(N-k)} = \frac{(R_0^2 - R_1^2)/(k-l)}{(1-R_0^2)/(N-k)} \\ &= \frac{(0.751-0.682)/(6-4)}{(1-0.751)/(30-6)} = 3.325, \end{aligned}$$

vilket ej överstiger $F_{0.05}(2, 24) = 3.40$. Vi kan alltså inte förkasta nollhypotesen på nivån 5%.

d) Om vi utgår från medelkvadratsumman av residualerna från grundmodellen, får vi en väntevärdesriktig skattning av σ^2 oavsett om hypotesmodellen är sann eller inte. Det följer av (5) att denna skattning av feltermsvariansen är

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_i (Y_i - \hat{\mu}_i)^2}{N-k} = \frac{(1-R_0^2)\sum_i (Y_i - \bar{Y})^2}{N-k} = \frac{(1-R_0^2)\cdot \text{Kvs(Total)}}{N-k} \\ &= \frac{(1-0.751)\cdot 41.2}{30-6} = 0.4275. \end{aligned}$$

Motsvarande medelkvadratsumma av residualerna från hypotesmodellen är endast väntevärdesriktig om hypotesmodellen är sann.

Uppgift 3

a) Parametervektorn för regressionsmodellen är $\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2)^T$, och designmatrisen är

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X} \end{pmatrix},$$

där $\mathbf{1} = (1, 1, 1, 1, 1)^T$ och $\boldsymbol{X}$ består av de två x-kolumnerna av designmatrisen. Eftersom den första kolumnen i $\boldsymbol{A}$ är ortogonal mot de två andra kolumnerna (dvs de två förklarande variablerna är centrerade), följer att

$$\begin{aligned} \text{Var}\left[(\hat{\beta}_1, \hat{\beta}_2)\right] &= \sigma^2 (\boldsymbol{X}\boldsymbol{X}^T)^{-1} = \sigma^2 \boldsymbol{S}^{-1} = \sigma^2 \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}^{-1} \\ &= \sigma^2 \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \end{aligned}$$

där vi i andra steget införde $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X} = (s_{ij})_{i,j=1}^2$, och i sista steget utnyttjades ledningen. Speciellt ser vi från första diagonalelementet av högerledets matris, att $\text{Var}(\hat{\beta}_1) = 0.5\sigma^2$.

b) Vi följer ledningen och subtraherar bort termen $\beta_2 x_{2i}$ från observation Y_i. Det ger regressionsmodellen

$$Y_i' = Y_i - \beta_2 x_{2i} = \alpha + \beta_1 x_{1i} + \varepsilon_i$$

för $i = 1, \ldots, 5$. Detta är en enkel linjär regression, med designmatris

$$\boldsymbol{A}_0 = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X}_0 \end{pmatrix} \Rightarrow \boldsymbol{X}_0^T \boldsymbol{X}_0 = 4,$$

där $\boldsymbol{X}_0$ är den andra kolumnen av $\boldsymbol{A}_0$. Av detta följer att

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 (\boldsymbol{X}_0^T \boldsymbol{X}_0)^{-1} = 0.25\sigma^2.$$

c) Variansinflationsfaktorn (VIF) för skattningen av β_1 anger hur mycket dess varians ökar på grund av att β_2 måste skattas. Med andra ord anger VIF hur mycket $\mathrm{Var}(\hat{\beta}_1)$ ökar på grund av att β_2 måste skattas. Det följer av a) och b) att

$$\mathrm{VIF} = \frac{0.5 \cdot \sigma^2}{0.25 \cdot \sigma^2} = 2.$$

Vi kan också utnyttja definitionen av variationsinflationsfaktorn. Enligt formel (3.43) i kompendiet (med beteckningen s_{11} i stället för s_{11}^2) fås

$$\mathrm{VIF} = s_{11} \cdot (\boldsymbol{S}^{-1})_{11} = 4 \cdot 0.5 = 2.$$

Ett tredje alternativ är att utnyttja formeln

$$\mathrm{VIF} = \frac{1}{1 - R^2}, \tag{6}$$

där R^2 är förklaringsgraden av $\{x_{1i}\}$, i en enkel linjär regression där $\{x_{2i}\}$ utgör den enda oberoende variabeln. För att räkna ut R^2 ansätter vi därför

$$x_{1i} = a + b x_{2i} + \epsilon_i, \quad i = 1, \ldots, 5.$$

Eftersom $\{x_{2i}\}$ redan är centrerade ($\bar{x}_2 = 0$), följer att minsta kvadratskattningen av lutningparametern är

$$\hat{b} = \frac{\sum_{i=1}^5 x_{2i} x_{1i}}{\sum_{i=1}^5 x_{2i}^2} = \frac{s_{12}}{s_{22}} = \frac{2}{2} = 1.$$

Det ger

$$\begin{aligned} R^2 &= \frac{\mathrm{Kvs(Regression)}}{\mathrm{Kvs(Total)}} = \frac{\sum_{i=1}^5 (\hat{b} \cdot (x_{2i} - \bar{x}_2))^2}{\sum_{i=1}^5 (x_{1i} - \bar{x}_1)^2} = \frac{\sum_{i=1}^5 (\hat{b} \cdot x_{2i})^2}{\sum_{i=1}^5 x_{1i}^2} \\ &= \frac{\hat{b}^2 s_{22}}{s_{11}} = \frac{1^2 \cdot 2}{4} = 0.5. \end{aligned}$$

Insättning av detta värde i (6) ger VIF = 2.

Uppgift 4

a) Eftersom

$$s_{ij}^2 = \frac{1}{2}\sum_{k=1}^{3}(Y_{ijk} - \bar{Y}_{ij\cdot})^2,$$

och antalet frihetsgrader för variationskällan Inom celler är $3 \cdot 3 \cdot 2 = 18$, så följer att

$$\begin{aligned}\hat{\sigma}^2 &= \text{Mkvs(Inom celler)} = \tfrac{\text{Kvs(Inom celler)}}{18} \\ &= \tfrac{1}{18}\textstyle\sum_{i,j,k=1}^{3}(Y_{ijk} - \bar{Y}_{ij\cdot})^2 = \tfrac{1}{9}\sum_{i,j=1}^{3} s_{ij}^2 = 0.52.\end{aligned}$$

b) Vi börjar med att räkna ut en skattning

$$\hat{\gamma}_{ij} = \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}$$

av sampspelstermen γ_{ij} för alla kombinationer i, j av träd och säsong, se tabellen nedan.

Värden på $\hat{\gamma}_{ij}$:

	$j=1$	$j=2$	$j=3$
$i=1$	0.2	0.4	-0.6
$i=2$	-0.8	-1.0	1.8
$i=3$	0.6	0.6	-1.2

Eftersom antal frihetsgrader för samspel är $(3-1)(3-1) = 4$, så ges dess medelkvadratsumma av

$$\text{Mkvs(Samspel)} = \frac{\text{Kvs(Samspel)}}{4} = \frac{1}{4}\sum_{i,j,k=1}^{3}\hat{\gamma}_{ij}^2 = \frac{3}{4}\sum_{i,j=1}^{3}\hat{\gamma}_{ij}^2 = \frac{3 \cdot 7.6}{4} = 5.7.$$

Under nollhypotesen $H_0 : \sigma_\gamma^2 = 0$ så är kvoten av medelkvadratsummorna i b) och a) F-fördelad. Eftersom dess värde

$$\text{F-kvot} = \frac{\text{Mkvs(Samspel)}}{\text{Mkvs(Inom celler)}} = \frac{5.7}{0.52} = 10.96$$

överstiger $F_{0.05}(4, 18) = 2.93$ så är samspelet signifikant.

c) Då vi har $n = 3$ replikat per cell, följer från formelsamlingen att

$$E\left[\text{Mkvs(Samspel)}\right] = \sigma^2 + 3\sigma_\gamma^2.$$

Genom att utnyttja resultaten i a) och b), kan vi därför skatta samspelsvariansen med

$$\hat{\sigma}_\gamma^2 = \frac{\text{Mkvs(Samspel)} - \hat{\sigma}^2}{3} = \frac{5.7 - 0.52}{3} = 1.727.$$

Uppgift 5

a) En ARMA$(1,1)$-process är stationär om och endast om $|\phi| < 1$, oavsett värdet på θ.

b) Vi börjar med att använda definitionen av en ARMA$(1,1)$-process för att beräkna

$$\begin{aligned} \text{Cov}(X_t, \varepsilon_t) &= \text{Cov}(\phi X_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}, \varepsilon_t) \\ &= \phi\text{Cov}(X_{t-1}, \varepsilon_t) + \text{Var}(\varepsilon_t) - \theta\text{Cov}(\varepsilon_{t-1}, \varepsilon_t) \\ &= \phi \cdot 0 + \sigma_\varepsilon^2 - \theta \cdot 0 \\ &= \sigma_\varepsilon^2, \end{aligned} \tag{7}$$

där vi i tredje ledet utnyttjade ledningen. Därefter beräknar vi

$$\begin{aligned} \gamma_0 &= \text{Var}(X_t) \\ &= \text{Var}(\phi X_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}) \\ &= \phi^2\text{Var}(X_{t-1}) + \text{Var}(\varepsilon_t) + \theta^2\text{Var}(\varepsilon_{t-1}) + 2\phi\text{Cov}(X_{t-1}, \varepsilon_t) \\ &\quad -2\phi\theta\text{Cov}(X_{t-1}, \varepsilon_{t-1}) \\ &= \phi^2\gamma_0 + \sigma_\varepsilon^2 + \theta^2\sigma_\varepsilon^2 + 2\phi \cdot 0 - 2\phi\theta\sigma_\varepsilon^2 \\ &= \phi^2\gamma_0 + (1+\theta^2)\sigma_\varepsilon^2 - 2\phi\theta\sigma_\varepsilon^2, \end{aligned} \tag{8}$$

där vi i fjärde ledet utnyttjade ledningen och (7). Ekvation (8) är linjär i γ_0. Genom att lösa ut γ_0 erhålls

$$\gamma_0 = \sigma_\varepsilon^2 \cdot \frac{1+\theta^2-2\phi\theta}{1-\phi^2}. \tag{9}$$

Slutligen fås kovariansfunktionens värde för $k = 1$ enligt

$$\begin{aligned} \gamma_1 &= \text{Cov}(X_t, X_{t+1}) \\ &= \text{Cov}(X_t, \phi X_t + \varepsilon_{t+1} - \theta\varepsilon_t) \\ &= \phi\text{Var}(X_t) + \text{Cov}(X_t, \varepsilon_{t+1}) - \theta\text{Cov}(X_t, \varepsilon_t) \\ &= \phi\gamma_0 + 0 - \theta \cdot \sigma_\varepsilon^2, \end{aligned} \tag{10}$$

där vi i sista ledet utnyttjade ledningen och (7). Insättning av (9) i (10) ger

$$\gamma_1 = \sigma_\varepsilon^2 \left(\phi\frac{1+\theta^2-2\phi\theta}{1-\phi^2} - \theta \right) = \sigma_\varepsilon^2 \cdot \frac{(\phi-\theta)(1-\phi\theta)}{1-\phi^2}. \tag{11}$$

c) Givet ϕ ser vi från (11) att $\gamma_1 = 0$ för $\theta = \phi$ och $\theta = 1/\phi$. Genom att använda bakåtoperatorn B ser man att

$$(1-\phi B)X_t = (1-\theta B)\varepsilon_t \iff X_t = \frac{1-\theta B}{1-\phi B}\varepsilon_t.$$

Om $\theta = \phi$ följer att täljare och nämnare i den sista kvoten tar ut varandra, så att $X_t = \varepsilon_t$. Så är inte fallet om $\theta = 1/\phi$.