STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

11 december 2024 14–19

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

Uppgift 1

a) Eftersom

$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{1}{9}(2015 + 2016 + \ldots + 2023) = 2019$$

och

$$\begin{aligned} \textstyle\sum_{i=1}(x_i - \bar{x})^2 &= [(2015 - 2019)^2 + \ldots + (2023 - 2019)^2] \\ &= (16 + 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16) \\ &= 60, \end{aligned}$$

så följer att minsta kvadrat-skattningen av lutningsparametern β ges av

$$\hat{\beta} = \frac{\sum_{i=1}^{9}(x_i - \bar{x})y_i}{\sum_{i=1}^{9}(x_i - \bar{x})^2} = \frac{24.1}{60} = 0.4017.$$

b) Kvadratsumman för variationskällan Residual ges av

$$\text{Kvs(Residual)} = \text{Kvs(Total)} - \text{Kvs(Regression)} = 10.12 - 9.68 = 0.44,$$

och den har $N - 2 = 9 - 2 = 7 = f$ frihetsgrader. Därur följer att

$$\hat{\sigma}^2 = \text{Mkvs(Residual)} = \frac{\text{Kvs(Residual)}}{7} = 0.0628$$

är en väntevärdesriktig skattning av feltermernas varians.
c) Skattningen av lutningsparametern har en varians

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{9}(x_i - \bar{x})^2} = \frac{\sigma^2}{60},$$

och därför är dess medelfel

$$d = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})} = \sqrt{\frac{\hat{\sigma}^2}{60}} = \sqrt{\frac{0.0628}{60}} = 0.0324.$$

Således får vi ett 95% konfidensintervall

$$\begin{aligned} I_\beta &= (\hat{\beta} - t_{0.025}(7)d, \hat{\beta} + t_{0.025}(7)d) \\ &= (0.4017 - 2.365 \cdot 0.0324, 0.4017 + 2.365 \cdot 0.0324) \\ &= (0.325, 0.478) \end{aligned}$$

för den årliga ökningstakten av BNP-tillväxten, där $t_{0.025}(7) = \sqrt{F_{0.05}(1,7)}$ fås ur tabell. Eftersom detta intervall inte innehåller 0 så har ökningen av BNP-tillväxten varit signifikant (på nivån 5%) mellan åren 2015 och 2023.

Uppgift 2

a) Vi kan skriva den multipla linjära regressionsmodellen på matrisform enligt $\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon}$, eller med utskrivna element i de ingående vektorerna och matriserna, som

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{25} \\ Y_{26} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & 0 & 0 \\ 1 & x_{12} & x_{22} & 0 & 0 \\ 1 & x_{13} & x_{23} & 0 & 0 \\ 1 & x_{14} & x_{24} & 0 & 0 \\ 1 & x_{15} & x_{25} & 0 & 0 \\ 1 & x_{16} & x_{26} & 0 & 0 \\ 1 & 0 & 0 & x_{11} & x_{21} \\ 1 & 0 & 0 & x_{12} & x_{22} \\ 1 & 0 & 0 & x_{13} & x_{23} \\ 1 & 0 & 0 & x_{14} & x_{24} \\ 1 & 0 & 0 & x_{15} & x_{25} \\ 1 & 0 & 0 & x_{16} & x_{26} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \\ \varepsilon_{26} \end{pmatrix}.$$

b) Nollhypotesen H_0 svarar mot att införa gemensamma effektparametrar $\beta_1 = \beta_{11} = \beta_{21}$ och $\beta_2 = \beta_{12} = \beta_{22}$ för båda maskinernas förmåga att eliminera bakterie 1 respektive bakterie 2. Vi får en parametervektor $\boldsymbol{\lambda} = (\alpha, \beta_1, \beta_2)^T$ med tre parametrar under H_0, vilket svarar mot att den ursprungliga parametervektorn måste uppfylla

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \boldsymbol{B\lambda}.$$

c) Eftersom antalet parametrar för grundmodellen och hypotesmodellen är $k = 5$ respektive $l = 3$, är antal frihetsgrader för variationskällan Avvikelse från H_0 lika med $k - l = 5 - 3 = 2$. Den andra variationskällan, Residual, har $N - k = 12 - 5 = 7$ frihetsgrader. Vi kan därför testa nollhypotesen med en

$$\text{F-kvot} = \frac{\text{Mkvs(Avv från } H_0)}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Avv från } H_0)/2}{\text{Kvs(Residual)}/7} = \frac{8.00/2}{8.14/7} = 3.44,$$

vars värde understiger $F_{0.05}(2, 7) = 4.7$. Vi kan därför inte förkasta nollhypotesen att de två vattenrenarna är likvärdiga när det gäller att filtrera bort de två typerna av bakterier.

Uppgift 3

a) För att skatta σ_ϵ^2 använder vi medelkvadratsumman för variationskällan Inom stavar. Eftersom antalet frihetsgrader för denna variationskälla är $4(5-1) = 16$, så följer att

$$\begin{aligned} \hat{\sigma}_\epsilon^2 &= \tfrac{1}{16}\text{Kvs(Inom stavar)} \\ &= \tfrac{1}{16}\textstyle\sum_{i=1}^{4}\sum_{j=1}^{5}(Y_{ij} - \bar{Y}_{i\cdot})^2 \\ &= \tfrac{1}{16}\textstyle\sum_{i=1}^{4} 4s_i^2 \\ &= \tfrac{1}{4}\textstyle\sum_{i=1}^{4} s_i^2 \\ &= \tfrac{1}{4}(0.10 + 0.14 + 0.08 + 0.12) \\ &= 0.11, \end{aligned} \tag{1}$$

där vi i tredje ledet införde definitionen av stickprovsvariansen $s_i^2 = \sum_{j=1}^{5}(Y_{ij} - \bar{Y}_{i\cdot})^2/(5-1)$ för de uppmätta strålningsvärdena hos stav i, för $i = 1, 2, 3, 4$.

b) Radmedelvärdena är oberoende och normalfördelade

$$\bar{Y}_{i\cdot} = \mu + \delta_i + \bar{\varepsilon}_{i\cdot} \sim N(\mu, \sigma_\delta^2 + \frac{\sigma_\epsilon^2}{5}).$$

Vi skattar sedan variansen $V = \sigma_\delta^2 + \sigma_\epsilon^2/5$ för radmedelvärdena enligt

$$\begin{aligned} \hat{V} &= \hat{\sigma}_\delta^2 + \tfrac{\hat{\sigma}_\epsilon^2}{5} \\ &= \tfrac{1}{4-1}\textstyle\sum_{i=1}^{4}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= \tfrac{1}{3}[(80.0 - 80.0)^2 + (82.0 - 80.0)^2 + (78.5 - 80.0)^2 + (79.5 - 80.0)^2] \\ &= 2.1667, \end{aligned} \tag{2}$$

som alternativt kan beräknas direkt från den i tabellen givna stickprovsstandardavvikelsen av radmedelvärdena, enligt $\hat{V} = \text{Std}^2$. Genom att kombinera (1) och (2), får vi slutligen en väntevärdesriktig skattning

$$\hat{\sigma}_\delta^2 = 2.1667 - \frac{0.11}{5} = 2.145 \tag{3}$$

av σ_δ^2. Vi noterar att eftersom $\hat{\sigma}_\epsilon^2$ är mycket mindre än $\hat{\sigma}_\delta^2$, hade det förmodligen räckt att göra en mätning per stav.

En alternativ lösning är att utnyttja

$$E[\text{Mkvs(Mellan stavar)}] = 5\sigma_\delta^2 + \sigma_\varepsilon^2$$

från formelsamlingen, vilket medför

$$\begin{aligned} 5\hat{\sigma}_\delta^2 + \hat{\sigma}_\epsilon^2 &= \text{Mkvs(Mellan stavar)} \\ &= \text{Kvs(Mellan stavar)}/3 \\ &= \textstyle\sum_{i=1}^{4}\sum_{j=1}^{5}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/3 \\ &= 5 \cdot \textstyle\sum_{i=1}^{4}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/3 \\ &= 5 \cdot 2.1667, \end{aligned}$$

eftersom varitionskällan Mellan stavar har $4 - 1 = 3$ frihetsgrader. Därefter utnyttjar vi (1) för att komma fram till (3).

c) Som skattning av den förväntade strålningsmängden μ från en slumpmässigt vald stav, väljer vi

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} = \frac{1}{4}\sum_{i=1}^{4}\bar{Y}_{i\cdot} = \frac{1}{4}(80.0 + 82.0 + 78.5 + 79.5) = 80.0.$$

Eftersom

$$\text{Var}(\hat{\mu}) = \frac{V}{4},$$

så får vi ett medelfel

$$d = \sqrt{\widehat{\text{Var}}(\hat{\mu})} = \sqrt{\frac{\hat{V}}{4}} = \frac{\sqrt{2.1667}}{2} = \frac{1.472}{2} = 0.736,$$

och ett 95% konfidensintervall

$$\begin{aligned} (\hat{\mu} - t_{0.025}(3)d, \hat{\mu} + t_{0.025}(3)d) &= (80.0 - 3.182 \cdot 0.736, 80.0 + 3.182 \cdot 0.736) \\ &= (77.66, 82.34) \end{aligned}$$

för μ, där värdet på $t_{0.025}(3) = \sqrt{F_{0.05}(1,3)}$ kan fås ur tabell.

Uppgift 4

a) Vi börjar med att beräkna

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

och

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

För $k \geq 2$ gäller att $X_t = \varepsilon_t - \theta_{t-1}\varepsilon_{t-1}$ och $X_{t+k} = \varepsilon_{t+k} - \theta\varepsilon_{t+k-1}$ är funktioner av olika feltermer. Därför är X_t oberoende av X_{t+k} då $k \geq 2$, vilket medför att $\gamma_k = 0$ för $k \geq 2$.

b) Det följer av a) och stationäriteten hos $\{X_t\}$ att

$$\rho_k = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\sqrt{\mathrm{Var}(X_t)}\sqrt{\mathrm{Var}(X_{t+k})}} = \frac{\gamma_k}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_k}{\gamma_0} = \begin{cases} 1, & k = 0, \\ -\theta/(1+\theta^2), & k = 1, \\ 0, & k \geq 2. \end{cases}$$

c) Korrelationsfunktionen ρ_k är identisk för två MA(1)-processer med parametrar θ och $1/\theta$. Det inses av att $\rho_k = 0$ för $k \geq 2$ för båda MA(1)-processerna, och att

$$\rho_1 = -\frac{\theta}{1+\theta^2} = -\frac{1/\theta}{1+(1/\theta)^2}$$

har samma värde för de två processerna. För $\theta = 1$ är processerna identiska, men inte för $\theta \neq 1$. Det följer av att MA(1)-processen är inverterbar endast då $|\theta| < 1$, det vill säga att feltermen ε_t då kan skriva som en konvergerande linjärkombination

$$\varepsilon_t = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \theta^3 X_{t-3} + \ldots$$

av alla värden på MA(1)-processen fram till tiden t.

Uppgift 5

a) Låt $\boldsymbol{Y} = (Y_1, \ldots, Y_6)^T$ vara observationsvektorn och $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{j6})^T$ vektorn med värden på kovariat $j = 1, 2$. För delmodellen med endast kovariat 1 så ges minsta kvadrat-skattningen av β_1, av

$$\begin{aligned}\hat{\beta}_1 &= \boldsymbol{x}_1\boldsymbol{Y}/(\boldsymbol{x}_1\boldsymbol{x}_1^T) \\ &= \tfrac{0(-0.1)+0(0.1)+1\cdot 2+(-1)0+(-1)(-2)+1\cdot 0}{0^2+0^2+1^2+(-1)^2+(-1)^2+1^2} \\ &= 4/4 \\ &= 1.\end{aligned}$$

Skattningen av feltermsvariansen blir

$$\begin{aligned}\hat{\sigma}^2 &= \textstyle\sum_{i=1}^6 (Y_i - \hat{\beta}_1 x_{1i})^2/(6-1) \\ &= \textstyle\sum_{i=1}^6 (Y_i - x_{1i})^2/5 \\ &= [(-0.1)^2 + 0.1^2 + 1^2 + 1^2 + (-1)^2 + (-1)^2]/5 \\ &= 4.02/5 \\ &= 0.804.\end{aligned}$$

För att testa nollhypotesen att ingen kovariat ingår i modellen mot delmodellen att bara kovariat 1 ingår, får vi alltså en

$$\begin{aligned} \text{F-kvot} &= \text{Mkvs(Regression)}/\hat{\sigma}^2 \\ &= \text{Kvs(Regression)}/\hat{\sigma}^2 \\ &= \hat{\beta}_1^2 \textstyle\sum_i x_{1i}^2/\hat{\sigma}^2 \\ &= 4\hat{\beta}_1^2/\hat{\sigma}^2 \\ &= 4/0.804 \\ &= 4.975, \end{aligned}$$

där vi i andra ledet utnyttjade att variationskällan Regression har 1 frihetsgrad. I tredje ledet utnyttjade vi att skattningarna av $\boldsymbol{\mu} = E(\boldsymbol{Y})$ är $\hat{\hat{\boldsymbol{\mu}}} = (0, \ldots, 0)^T$ under nollhypotesen och $\hat{\boldsymbol{\mu}} = \hat{\beta}_1 \boldsymbol{x}_1$ under grundmodellen. Således är kvadratsumman för regression samma sak som kvadratsumman för avvikelse från nollhypotes; $\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 = \hat{\beta}_1^2 \sum_i x_{1i}^2$. Eftersom F-kvoten ovan inte överstiger tröskelvärdet $F_{0.05}(1, 5) = 6.6$, så kan vi inte förkasta nollhypotesen att inte ta med kovariat 1 i modellen, på nivån 5%.

Om vi istället testar samma nollhypotes mot modellen med kovariat 2, så inser vi av symmetriskäl att $\hat{\beta}_2 = 1$, $\hat{\sigma}^2 = 0.804$ och F-kvot $= 4\hat{\beta}_2^2/\hat{\sigma}^2 = 4.975 < F_{0.05}(1, 5)$. För detta test förkastas alltså inte heller nollhypotesen på nivån 5%.

Slutsatsen blir att FS-schemat stannar efter första steget, och den valda modellen är den som inte innehåller några kovariater. Anledningen är att den kovariat som inte tas med i respektive test fångas upp av residualerna. Därför dränks den kovariat man vill testa i residualernas brus, så att den inte får en signifikant inverkan på responsvariablerna.

b) Enligt BE-metoden börjar vi med den fulla modellen där båda kovariaterna ingår. Vi testar först nollhypotesen H_0 att bara kovariat 1 ingår mot den fulla modellen. Eftersom de två kolumnera i designmatrisen $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ för den fulla modellen är ortogonala ser man att skattningarna av β_1 och β_2 är desamma (=1) som i 5a), även för modellen där båda kovariaterna ingår. Eftesom skattningen av $\boldsymbol{\mu} = E(\boldsymbol{Y})$ är $\hat{\hat{\boldsymbol{\mu}}} = \hat{\beta}_1 \boldsymbol{x}_1 = \boldsymbol{x}_1$ för hypotesmodellen H_0, och $\hat{\boldsymbol{\mu}} = \boldsymbol{X}(\hat{\beta}_1, \hat{\beta}_2)^T = \boldsymbol{x}_1 + \boldsymbol{x}_2$ under grundmodellen med båda kovariaterna, fås att

$$\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 = \|\boldsymbol{x}_2\|^2 = 4,$$

medan

$$\begin{aligned} \hat{\sigma}^2 &= \textstyle\sum_{i=1}^6 (Y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2/(6-2) \\ &= \textstyle\sum_{i=1}^6 (Y_i - x_{1i} - x_{2i})^2/4 \\ &= [(-0.1)^2 + (0.1)^2 + 0^2 + 0^2 + 0^2 + 0^2]/4 \\ &= 0.005. \end{aligned}$$

Eftersom grundmodellen innehåller 1 parameter mer än hypotesmodellen ($k - l = 2 - 1 = 1$), ger det en

$$\text{F-kvot} = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/(k-l)}{\hat{\sigma}^2} = \frac{4}{0.005} = 800$$

som vida överstiger tröskelvärdet $F_{0.05}(1, 4) = 7.7$.

Av symmetriskäl får vi samma resultat då hypotesmodellen att bara kovariat 2 ingår i modellen testas mot den fulla modellen med båda kovariaterna, dvs en F-kvot $= 800 > F_{0.05}(1, 4)$.

Slutsatsen blir att BE-schemat stannar efter första steget, dvs att modellen med både kovariat 1 och 2 väljs.

c) Vi ser att BE-metoden ger ett rimligare resultat än FS. Man kan lätt visa (t ex genom att räkna ut förklaringsgraderna) att den fulla modellen fångar upp i stort sätt all variation i responsvariabeln, de två delmodellerna med en kovariat fångar upp hälften av denna variation, medan modellen utan kovariater inte fångar upp någonting av variationen i Y.