STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

20 oktober 2022 14–19

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

Uppgift 1

a) Eftersom $\bar{x} = 2016$ följer att $\sum_i (x_i - \bar{x})^2 = 2(1^2 + 2^2 + 3^2 + 4^2) = 60$. Det ger minsta kvadrat-skattningar

$$\begin{aligned} \hat{\tilde{\alpha}} &= \bar{y} = 26.9, \\ \hat{\beta} &= \textstyle\sum_i y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 = 27.3/60 = 0.455 \end{aligned}$$

av intercept och lutning. Det förväntade kilopriset 2022 är

$$\mu(2022) = \tilde{\alpha} + \beta(2022 - \bar{x}) = \tilde{\alpha} + 6\beta,$$

och det skattas med

$$\hat{\mu}(2022) = \hat{\tilde{\alpha}} + 6\hat{\beta} = 26.9 + 6 \cdot 0.455 = 29.63.$$

b) Eftersom vi har $N = 9$ observationer följer att

$$\begin{aligned} \mathrm{Var}(\hat{\mu}(2022)) &= \mathrm{Var}(\hat{\tilde{\alpha}}) + 6^2 \cdot \mathrm{Var}(\hat{\beta}) = \sigma^2 \left(\tfrac{1}{9} + \tfrac{6^2}{\sum_{i=1}^9 (x_i - \bar{x})^2} \right) \\ &= \sigma^2 \left(\tfrac{1}{9} + \tfrac{36}{60} \right) = 0.711\sigma^2, \end{aligned}$$

där felvariansen skattas med

$$\hat{\sigma}^2 = \frac{\mathrm{Kvs(Residual)}}{N - 2} = \frac{0.9274}{7} = 0.1325.$$

Det ger ett medelfel

$$d = \sqrt{\widehat{\mathrm{Var}}(\hat{\mu}(2022))} = \sqrt{0.711\hat{\sigma}^2} = \sqrt{0.711 \cdot 0.1325} = 0.3069.$$

c) Konfidensintervallet ges av

$$I = \hat{\mu}(2022) \pm t_{0.025}(N - 2)d = 29.63 \pm 2.3646 \cdot 0.3069 = (28.90, 30.36).$$

Uppgift 2

a) Matrisformuleringen av den linjära modellen $\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon}$ ges av

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{25} \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 & 0 \\ 1 & x_2 & 0 & 0 \\ 1 & x_3 & 0 & 0 \\ 1 & x_4 & 0 & 0 \\ 1 & x_5 & 0 & 0 \\ 0 & 0 & 1 & x_1 \\ 0 & 0 & 1 & x_2 \\ 0 & 0 & 1 & x_3 \\ 0 & 0 & 1 & x_4 \\ 0 & 0 & 1 & x_5 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \end{pmatrix}.$$

b) Låt $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ innehålla det gemensamma interceptet $\lambda_1 = \alpha_1 = \alpha_2$ och den gemensamma effektparametern $\lambda_2 = \beta_1 = \beta_2$ under nollhypotesen H_0. Denna nollhypotes svarar mot $\boldsymbol{\theta} = \boldsymbol{B\lambda}$, där

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

c) Eftersom antalet observationer $N = 10$, grundmodellen har $k = 4$ parametrar och hypotesmodellen $l = 2$ parametrar följer att

$$\begin{aligned} \text{Mkvs(Avv. från } H_0) &= \text{Kvs(Avv. från } H_0)/(k-l) = 6.8910/2 = 3.4455, \\ \text{Mkvs(Residual)} &= \text{Kvs(Residual)}/(N-k) = 1.1060/6 = 0.1843. \end{aligned}$$

Det ger en

$$\text{F-kvot} = \frac{\text{Mkvs(Avv. från } H_0)}{\text{Mkvs(Residual)}} = \frac{3.4455}{0.1843} = 18.69$$

som överstiger $F_{0.05}(k-l, N-k) = F_{0.05}(2,6) = 5.1$. Vi kan alltså förkasta nollhypotesen att medicinerna har samma effekt.

Uppgift 3

a) Antalet frihetsgrader för residualerna $Y_{ij} - \bar{Y}_{i\cdot}$ är $5(3-1) = 10$. Det ger en skattning

$$\begin{aligned} \hat{\sigma}^2_\varepsilon &= \text{Mvs(Inom block)} \\ &= \tfrac{1}{10}\text{Kvs(Inom block)} \\ &= \tfrac{1}{10}\textstyle\sum_{i=1}^{5}\sum_{j=1}^{3}(Y_{ij} - \bar{Y}_{i\cdot})^2 \\ &= \tfrac{1}{10}\textstyle\sum_{i=1}^{5} 2s_i^2 \\ &= \tfrac{1}{5}\textstyle\sum_{i=1}^{5} s_i^2 \\ &= \tfrac{1}{5}(0.035 + 0.031 + 0.020 + 0.046 + 0.023) \\ &= 0.031 \end{aligned}$$

av σ_ε^2. Eftersom radmedelvärdena

$$\bar{Y}_{i\cdot} = \mu + \delta_i + \bar{\varepsilon}_{i\cdot} \sim N(\mu, \sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{3})$$

är sinsemellan oberoende, kan vi använda deras stickprovsvarians för att skatta $\sigma_\delta^2 + \sigma_\varepsilon^2/3$:

$$\begin{aligned}\hat{\sigma}_\delta^2 + \tfrac{1}{3}\hat{\sigma}_\varepsilon^2 &= \tfrac{1}{5-1}\sum_{i=1}^{5}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= \tfrac{1}{4}\left[(3.1-2.9)^2 + (2.6-2.9)^2 + (2.8-2.9)^2 + (3.3-2.9)^2 \right. \\ &\quad \left. +(2.7-2.9)^2\right]. \\ &= 0.085.\end{aligned}$$

Det ger i sin tur en skattning

$$\hat{\sigma}_\delta^2 = 0.085 - \frac{0.031}{3} = 0.075$$

av σ_δ^2.

b) Minsta kvadrat-skattningen av μ är $\hat{\mu} = \bar{Y}_{\cdot\cdot} = 2.9$. Eftersom

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma_\delta^2}{5} + \frac{\sigma_\varepsilon^2}{15}$$

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

$$\begin{aligned} d &= \sqrt{\widehat{\mathrm{Var}}(\hat{\mu})} \\ &= \sqrt{\tfrac{1}{5}(\hat{\sigma}_\delta^2 + \tfrac{1}{3}\hat{\sigma}_\varepsilon^2)} \\ &= \sqrt{\tfrac{0.085}{5}} \\ &= 0.1304, \end{aligned}$$

och konfidensintervall

$$\begin{aligned} I &= \hat{\mu} \pm t_{0.025}(4)d \\ &= 2.9 \pm \sqrt{F_{0.05}(1,4)} \cdot 0.1304 \\ &= 2.9 \pm \sqrt{7.7} \cdot 0.1304 \\ &= (2.54, 3.26) \end{aligned}$$

för μ med konfidensgrad 95%.

Uppgift 4

a) AR(1)-processen är stationär då $|\phi| < 1$.

b) Sätt $X_t = Y_t - \mu$, så att

$$X_t = \phi X_{t-1} + \varepsilon_t. \tag{1}$$

Eftersom X_{t-1} beror av feltermer $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ så kommer X_{t-1} att vara oberoende av ε_t. Av detta och stationäriteten hos $\{X_t\}$ följer att

$$\begin{aligned} \sigma^2 &= \mathrm{Var}(Y_t) = \mathrm{Var}(X_t) = \mathrm{Var}(\phi X_{t-1} + \varepsilon_t) \\ &= \phi^2 \mathrm{Var}(X_{t-1}) + \mathrm{Var}(\varepsilon_t) = \phi^2\sigma^2 + \sigma_\varepsilon^2. \end{aligned}$$

Löser vi ut denna ekvation med avseende på σ^2 så erhålls

$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1-\phi^2}.$$

Vidare har vi, givet $k \geq 1$, att

$$\begin{aligned} \gamma_k &= \mathrm{Cov}(Y_t, Y_{t+k}) = \mathrm{Cov}(X_t, X_{t+k}) = \mathrm{Cov}(X_t, \phi X_{t+k-1} + \varepsilon_{t+k}) \\ &= \phi \mathrm{Cov}(X_t, X_{t+k-1}) + \mathrm{Cov}(X_t, \varepsilon_{t+k}) = \phi\gamma_{k-1}. \end{aligned} \tag{2}$$

Eftersom $\gamma_0 = \sigma^2$ så medför (2) (med induktionsbevis) att $\gamma_k = \gamma_0\phi^k$. Det ger

$$\rho_k = \frac{\mathrm{Cov}(Y_t, Y_{t+k})}{\sqrt{\mathrm{Var}(Y_t)}\sqrt{\mathrm{Var}(Y_{t+k})}} = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\sqrt{\mathrm{Var}(X_t)}\sqrt{\mathrm{Var}(X_{t+k})}} = \frac{\gamma_k}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_k}{\gamma_0} = \phi^k.$$

c) Med hjälp av ledningen kan vi skriva (den approximativa) log likelihooden som

$$l(\boldsymbol{\psi}) = l(\alpha, \phi, \sigma_\varepsilon^2) = C_1 - C_2 \sum_{t=2}^{T} (y_t - \alpha - \phi y_{t-1})^2, \tag{3}$$

där $C_1 = -(T-1)\log(\sqrt{2\pi}\sigma_\varepsilon)$ och $C_2 = (2\sigma_\varepsilon^2)^{-1}$ är positiva och beror av σ_ε^2. Oavsett värdet på σ_ε^2 så maximeras $l(\boldsymbol{\psi})$ genom att minimera kvadratsumman i (3). Detta är ekvivalent med att skatta (icke-centrerat) intercept α och lutning ϕ med minsta kvadratmetoden i en enkel linjär regression med data $\{(y_{t-1}, y_t)\}_{t=2}^T$, där $x_t = y_{t-1}$ och y_t svarar mot förklarande variabel och responsvariabel för observation $(x_t, y_t) = (y_{t-1}, y_t)$. Med $\bar{x} = \sum_{t=2}^T x_t/(T-1) = \sum_{t=1}^{T-1} y_t/(T-1)$ och $\bar{y} = \sum_{t=2}^T y_t/(T-1)$ får vi alltså

$$\begin{aligned} \hat{\phi} &= \textstyle\sum_{t=2}^T (y_{t-1} - \bar{x}) y_t / \sum_{t=2}^T (y_{t-1} - \bar{x})^2, \\ \hat{\alpha} &= \bar{y} - \hat{\phi}\bar{x}. \end{aligned}$$

Eftersom $\mu = \alpha/(1-\phi)$ ger det ML-skattningen

$$\hat{\mu} = \hat{\alpha}/(1-\hat{\phi}) = (\bar{y} - \hat{\phi}\bar{x})/(1-\hat{\phi}) \approx \bar{y}$$

av μ, där vi i sista ledet utnyttjade att $\bar{x} \approx \bar{y}$ om T är stor.

Uppgift 5

a) Låt $\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$ vara minsta kvadrat-skattningen av parametervektorn $\boldsymbol{\theta}$. Härur följer att

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\hat{\boldsymbol{\theta}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}.$$

Eftersom detta samband ska gälla för alla $\boldsymbol{Y}$ så ges hattmatrisen av

$$\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T. \tag{4}$$

b) Vi kan skriva residualvektorn som

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{Y},$$

där $\boldsymbol{I}_N$ är identitetsmatrisen av ordning N. Av detta, och det faktum att $\mathrm{Var}(\boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_N$, följer att residualvektorn har kovariansmatrisen

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{e}) &= (\boldsymbol{I}_N - \boldsymbol{H})\mathrm{Var}(\boldsymbol{Y})(\boldsymbol{I}_N - \boldsymbol{H})^T \\
&= \sigma^2(\boldsymbol{I}_N - \boldsymbol{H})(\boldsymbol{I}_N - \boldsymbol{H})^T \\
&= \sigma^2(\boldsymbol{I}_N - \boldsymbol{H}^T - \boldsymbol{H} + \boldsymbol{H}\boldsymbol{H}^T) \\
&= \sigma^2(\boldsymbol{I}_N - \boldsymbol{H} - \boldsymbol{H} + \boldsymbol{H}) \\
&= \sigma^2(\boldsymbol{I}_N - \boldsymbol{H}).
\end{aligned} \tag{5}$$

I det fjärde steget av (5) utnyttjades att hattmatrisen är symmetrisk ($\boldsymbol{H} = \boldsymbol{H}^T$) och idempotent ($\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$), vilket följer ur (4) och räkneregler för matriser. Vidare ges väntevärdet för residualvektorn av

$$E(\boldsymbol{e}) = E[(\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{Y}] = (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu} = \boldsymbol{0}, \tag{6}$$

där vi i näst sista steget utnyttjade att $\boldsymbol{H}\boldsymbol{\mu} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{\mu}$. Eftersom $\mathrm{Var}(e_i)$ fås ur i:te diagonalelementet av kovariansmatrisen i (5), och $E(e_i)$ från komponent i av nollvektorn i (6), följer att

$$E(e_i^2) = E(e_i)^2 + \mathrm{Var}(e_i) = 0^2 + \sigma^2(\boldsymbol{I}_N - \boldsymbol{H})_{ii} = \sigma^2(1 - h_{ii}). \tag{7}$$

c) Om den förklarande variabeln centreras i den enkla linjära regressionsmodellen fås

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

där $\bar{x} = \sum_{i=1}^N x_i/N$. Av detta följer att designmatrisen ges av

$$\boldsymbol{A} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_N - \bar{x} \end{pmatrix} \Longrightarrow \boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} N & 0 \\ 0 & \sum_{j=1}^N (x_j - \bar{x})^2 \end{pmatrix}.$$

Sedan använder vi (4) för att räkna ut hattmatrisen enligt

$$\boldsymbol{H} = \boldsymbol{A}\begin{pmatrix} N & 0 \\ 0 & \sum_{j=1}^N (x_j - \bar{x})^2 \end{pmatrix}^{-1}\boldsymbol{A}^T = \boldsymbol{A}\begin{pmatrix} \frac{1}{N} & 0 \\ 0 & \frac{1}{\sum_{j=1}^N (x_j - \bar{x})^2} \end{pmatrix}\boldsymbol{A}^T.$$

Speciellt ges hattmatrisens i:te diagonalelement av

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{N}(x_j - \bar{x})^2}. \tag{8}$$

En inflytelserik punkt (x_i, Y_i) drar i större utsträckning till sig den skattade regressionslinjen, så att residualen e_i blir liten. Ju större värdet på h_{ii} är, desto mindre blir variansen för denna residual enligt (7). Det följer alltså av (8) att en observationspunkt kan förväntas vara mer inflytelserik ju längre dess förklarande variabel x_i ligger ifrån centrum $\bar{x}$ i förhållande till övriga observationer. Detta svarar mot den så kallade hävstångseffekten (leverage points).