STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

25 oktober 2024 14–19

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

Uppgift 1

a) Eftersom antal observationer $N = 6$, så ges minsta kvadrat-skattningarna av

$$\begin{aligned} \hat{\tilde{\alpha}} &= \textstyle\sum_i y_i/6 = 26.7008/6 = 4.450, \\ \hat{\beta} &= \textstyle\sum_i (x_i - \bar{x}) y_i / \sum_i (x_i - \bar{x})^2 = 10.9916/9.9995 = 1.099. \end{aligned}$$

b) Kovariansmatrisen ges av

$$\mathrm{Var}(\hat{\tilde{\alpha}}, \hat{\beta}) = \sigma^2 \begin{pmatrix} 1/6 & 0 \\ 0 & 1/\sum_i (x_i - \bar{x})^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 0.1667 & 0 \\ 0 & 0.1000 \end{pmatrix}. \quad (1)$$

c) Eftesom två regressionsparametrar ingår i modellen, så är antal frihetsgrader för Residual lika med $N - 2 = 6 - 2 = 4$. Det ger en skattning

$$\hat{\sigma}^2 = \mathrm{Mkvs(Residual)} = \frac{\mathrm{Kvs(Residual)}}{4} = \frac{0.04757}{4} = 0.001189,$$

av feltermsvariansen.

d) Låt $x_0 = \log(1609)$ vara den logaritmerade löpsträckan svarande mot en engelsk mil. Vi börjar med att ange en punktskattning av $\mu(x_0)$. Den ges av

$$\begin{aligned} \hat{\mu}(x_0) &= \hat{\tilde{\alpha}} + \hat{\beta}(x_0 - \bar{x}) \\ &= 4.450 + 1.099 \cdot (\log(1609) - 6.4017) \\ &= 5.529. \end{aligned}$$

För att bestämma medelfelet för denna skattning så noterar vi först, med hjälp av (1), att

$$\begin{aligned} \mathrm{Var}[\hat{\mu}(x_0)] &= \mathrm{Var}[\hat{\tilde{\alpha}} + \hat{\beta}(x_0 - \bar{x})] \\ &= \mathrm{Var}(\hat{\tilde{\alpha}}) + (x_0 - \bar{x})^2 \mathrm{Var}(\hat{\beta}) \\ &= \sigma^2 (\tfrac{1}{6} + \tfrac{(\log(1609) - 6.4017)^2}{9.9995}) \\ &= 0.2630 \cdot \sigma^2. \end{aligned}$$

Det ger ett medelfel

$$d = \sqrt{\widehat{\mathrm{Var}}[\hat{\mu}(x_0)]} = \sqrt{0.2630} \cdot \hat{\sigma} = 0.0177,$$

och ett 95% konfidensintervall

$$\begin{aligned} \hat{\mu}(x_0) \pm t_{0.025}(4)d &= (5.529 - 2.776 \cdot 0.0177, 5.529 + 2.776 \cdot 0.0177) \\ &= (5.480, 5.578) \end{aligned} \tag{2}$$

för $\mu(x_0)$, där t-kvantilen $t_{0.025}(4) = \sqrt{F_{0.05}(1,4)} = 2.776$ kan fås ur tabell. Eftersom $t = \exp(\mu(x_0))$ är en monoton transformation av $\mu(x_0)$, så får vi slutligen ett konfidensintervall

$$\begin{aligned} (\exp(5.4801), \exp(5.5783)) &= (239.9, 264.6) \\ &= (3 \text{ min } 59.5 \text{ s}, 4 \text{ min } 24.6 \text{ s}), \end{aligned}$$

för t, genom att applicera samma transformation på intervallet (2). (Notera att den undre gränsen är endast något högre än Lasses rekord på 1500 m. Det beror bland annat på att antal observationer är få, och eftersom Lasse är något sämre på medeldistans underksattar modellen hans löprekord på en engelsk mile. Observera även att *prediktionsintervallet* för en engelsk mile kommer att vara ännu vidare.)

Uppgift 2

a) De fyra mätningarna kan sammanfattas enligt

$$\begin{aligned} Y_1 &= \alpha + \beta_1 + \beta_2 + \varepsilon_1, \\ Y_2 &= \alpha - \beta_1 - \beta_2 + \varepsilon_2, \\ Y_3 &= \alpha + \beta_1 - \beta_2 + \varepsilon_3, \\ Y_4 &= \alpha - \beta_1 + \beta_2 + \varepsilon_4. \end{aligned}$$

Det svarar mot en multipel linjär regressionsmodell med designmatris

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

b) Man ser att $\boldsymbol{A}$ har ortogonala kolumner, så att

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} = 4\boldsymbol{I}_3.$$

Det ger minsta kvadrat-skattningar

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \tfrac{1}{4}\boldsymbol{A}^T\boldsymbol{Y} \\ &= \begin{pmatrix} (Y_1+Y_2+Y_3+Y_4)/4 \\ (Y_1-Y_2+Y_3-Y_4)/4 \\ (Y_1-Y_2-Y_3+Y_4)/4 \end{pmatrix} = \begin{pmatrix} 0.15 \\ 2.15 \\ 2.85 \end{pmatrix} \end{aligned} \tag{3}$$

av de tre regressionsparametrarna.

c) Eftersom $\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \sigma^2\boldsymbol{I}_3/4$, så följer att skattningarna av de tre parametrarna är oberoende (och normalfördelade) med samma varians

$$\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{4}.$$

Alternativt erhålls detta direkt från (3).

d) För att skatta medelfelet för $\hat{\beta}_1$ (och de två andra skattningarna), så börjar vi med att skatta σ^2. Detta kräver i sin tur en skattning

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\hat{\boldsymbol{\theta}} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 0.15 \\ 2.15 \\ 2.85 \end{pmatrix} = \begin{pmatrix} 5.15 \\ -4.85 \\ -0.55 \\ 0.85 \end{pmatrix}$$

av väntevärdet $\boldsymbol{\mu} = E(\boldsymbol{Y})$ för observationsvektorn $\boldsymbol{Y}$. Eftersom vi bara har $4-3=1$ frihetsgrad för att skatta feltermernas varians, så följer att denna skattning ges av residualkvadratsumman

$$\begin{aligned} \hat{\sigma}^2 &= \text{Kvs(Residual)} \\ &= \textstyle\sum_{i=1}^4 (Y_i - \hat{\mu}_i)^2 \\ &= (5.2-5.15)^2 + (-4.8-(-4.85))^2 + (-0.6-(-0.55))^2 + (0.8-0.85)^2 \\ &= 4\cdot 0.05^2 \\ &= 0.01. \end{aligned}$$

Med hjälp av c) och skattningen av σ^2 får vi ett medelfel

$$d = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2}{4}} = \frac{\hat{\sigma}}{2} = 0.05.$$

På grund av normalfördelningsantagandet får vi slutligen ett konfidensintervall

$$\begin{aligned} (\hat{\beta}_1 - t_{0.025}(1)d, \hat{\beta}_1 + t_{0.025}(1)d) &= (2.15 - 12.71\cdot 0.05, 2.15 + 12.71\cdot 0.05) \\ &= (1.52, 2.79), \end{aligned}$$

där $t_{0.025}(1) = \sqrt{F_{0.05}(1,1)} = 12.71$ avläses ur tabell. Notera att konfidensintervallet blir långt, eftersom man bara har en frihetsgrad för att skatta σ^2.

e) Om σ är känd så gäller att konfidensintervallen för β_1 och β_2 är $\hat{\beta}_1 \pm \lambda_{0.025}\sigma/2$ respektive $\hat{\beta}_2 \pm \lambda_{0.025}\sigma/2$, där $\lambda_{0.025}$ är 97.5%-kvantilen för en standard normalfördelning $N(0,1)$. I lösningen av c) såg vi att $\hat{\beta}_1$ och $\hat{\beta}_2$ är oberoende skattningar. Det två intervallet täcker därför över β_1 respektive β_2 oberoende av varandra, så att den simultana konfidensgraden för de två intervallen blir $0.95^2 = 0.9025$. (Detta gäller *inte* om σ är okänd, eftersom de två konfidensintervallen $\hat{\beta}_j \pm t_{0.025}(1)d$ då innehåller samma medelfel d.)

Uppgift 3

a) Eftersom vi har en tvåsidig variansanalys typ I med sampspel, bör varje cell i, j ha en egen parameter $E(Y_{ijk}) = \mu_{ij}$. Det ger totalt $3 \times 5 = 15$ parametrar, intercept och 14 regressionsparametrar (effektparametrar). I den aktuella parametriseringen $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ har vi 1 parameter för den genomsnittliga kundtrivseln μ hos alla företag (inteceptet) och 23 regressionsparametrar (3 parametrar α_i för företagsstorlekens inverkan, 5 parametrar β_j för företagsbranschens betydelse, och $3 \times 5 = 15$ samspelsparametrar γ_{ij}). Således har vi $23 - 14 = 9$ för många regressionsparametrar. Vi inför därför 9 oberoende linjära restriktioner på dessa 23 parametrar, nämligen

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 &= 0, \\ \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 &= 0, \\ \textstyle\sum_{j=1}^{5} \gamma_{ij} &= 0, \quad i = 1, \ldots, 3, \\ \textstyle\sum_{i=1}^{3} \gamma_{ij} &= 0, \quad j = 1, \ldots, 4. \end{aligned}$$

Notera att även $\sum_{i=1}^{3} \gamma_{i5} = 0$, men denna linjära restriktion är linjärt beroende av de som angivits ovan. De kvarvarande 14 regressionsparametrarna svarar mot totalt $14 = (3-1) + (5-1) + (3-1)(5-1)$ frihetsgrader för variationskällorna Storlek, Bransch och Samspel.

b) Antalet frihetsgrader för variationskällan Samspel är $(3-1)(5-1) = 8$, och för Inom celler är antalet frihetsgrader $3 \cdot 5(2-1) = 15$. För att testa nollhypotesen att det inte finns ett samspel mellan företagens storlek och branschtillhörighet vad gäller de anställdas trivsel, så bildar vi därför en

$$\text{F-kvot} = \frac{\text{Mkvs(Samspel)}}{\text{Mkvs(Inom celler)}} = \frac{\text{Kvs(Samspel)}/8}{\text{Kvs(Inom celler)}/15} = \frac{24.8/8}{30.75/15} = 1.51.$$

Då denna F-kvot inte överstiger tröskelvärdet $F_{0.05}(8, 15) = 2.6$, så kan vi inte på signifikansnivån 5% förkasta nollhypotesen.

c) Eftersom sampspel inte var signifikant i b), så inkluderar vi denna variationskälla bland residualerna, Det ger en kvadratsumma

$$\text{Kvs(Residual)} = \text{Kvs(Samspel)} + \text{Kvs(Inom celler)} = 24.80 + 30.75 = 55.55,$$

med 8+15=23 frihetsgrader. Variationskällan Storlek har $3 - 1 = 2$ frihetsgrader. Det innebär att vi kan testa nollhypotesen att företagens storlek inte

påverkar de anställdas trivsel, med en

$$\text{F-kvot} = \frac{\text{Mkvs(Storlek)}}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Storlek)}/2}{\text{Kvs(Residual)}/23} = \frac{22.60/2}{55.55/23} = 4.68.$$

Denna F-kvot överstiger $F_{0.05}(2, 23) = 3.4$, och således kan vi förkasta nollhypotesen på nivån 5%.

Uppgift 4

a) En AR(2)-process är stationär om båda rötterna till den karakteristiska ekvationen

$$x^2 - \phi_1 x - \phi_2 = 0 \tag{4}$$

ligger innanför enhetscirkeln i det komplexa talplanet.

b) Vi utnyttjar ledningen för att ge ett uttryck för kovariansfunktionen γ_k då $k = 1, 2$. Vi börjar med $k = 1$ och ser att

$$\begin{aligned} \gamma_1 &= \text{Cov}(X_t, X_{t+1}) = \text{Cov}(X_t, \phi_1 X_t + \phi_2 X_{t-1} + \varepsilon_{t+1}) \\ &= \phi_1 \text{Cov}(X_t, X_t) + \phi_2 \text{Cov}(X_t, X_{t-1}) + \text{Cov}(X_t, \varepsilon_{t+1}) \\ &= \phi_1 \gamma_0 + \phi_2 \gamma_1. \end{aligned} \tag{5}$$

Genom att lösa (5) med avseeende på γ_1 får vi

$$\gamma_1 = \phi_1 \gamma_0/(1 - \phi_2) \iff \rho_1 = \gamma_1/\gamma_0 = \phi_1/(1 - \phi_2). \tag{6}$$

Vi fortsätter med $k = 2$ och noterar att

$$\begin{aligned} \gamma_2 &= \text{Cov}(X_t, X_{t+2}) = \text{Cov}(X_t, \phi_1 X_{t+1} + \phi_2 X_t + \varepsilon_{t+2}) \\ &= \phi_1 \gamma_1 + \phi_2 \gamma_0 = \gamma_0(\phi_1 \rho_1 + \phi_2) \\ &= \gamma_0(\phi_1^2/(1 - \phi_2) + \phi_2), \end{aligned} \tag{7}$$

där vi i sista ledet utnyttjade (6). Division med γ_0 i båda leden av (7) leder slutligen till

$$\rho_2 = \gamma_2/\gamma_0 = \phi_1^2/(1 - \phi_2) + \phi_2.$$

c) Om x_1 och x_2 är de två rötterna till den karakteristiska ekvationen (4) så kan vi skriva

$$(x - x_1)(x - x_2) = x^2 - \phi_1 x - \phi_2 = 0.$$

Det innebär att $x_1 x_2 = -\phi_2$. Eftersom $|x_1| < 1$ och $|x_2| < 1$ måste gälla för en stationär AR(2)-process så följer att

$$|\phi_2| = |x_1||x_2| < 1 \cdot 1 = 1.$$

Anta vidare att $\phi_2 = -\phi_1^2/4$. Då har (4) en dubbelrot

$$x_1 = x_2 = \phi_1/2.$$

I detta fall är processen stationär ($|x_1| < 1$, $|x_2| < 1$) om och endast m $|\phi_1| < 2$.

Uppgift 5

a) Inför beteckningen PE_i för prediktionsfelet av observation Y_i. Vi har att

$$\begin{aligned} \text{PE}_i &= Y_i - \hat{\boldsymbol{\beta}}_{(-i)}^T \boldsymbol{x}_i \\ &= (\boldsymbol{\beta}^T \boldsymbol{x}_i + \varepsilon_i) - \hat{\boldsymbol{\beta}}_{(-i)}^T \boldsymbol{x}_i \\ &= \varepsilon_i - (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})^T \boldsymbol{x}_i. \end{aligned} \tag{8}$$

b) Notera först att eftersom $E(\hat{\boldsymbol{\beta}}_{(-i)}) = \boldsymbol{\beta}$, så följer att $E[(\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})^T \boldsymbol{x}_i] = [E(\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})]^T \boldsymbol{x}_i = \mathbf{0}^T \boldsymbol{x}_i = 0$, och därmed

$$E(\text{PE}_i) = E(\varepsilon_i) - E[(\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})^T \boldsymbol{x}_i] = 0 - 0 = 0. \tag{9}$$

Eftersom ε_i och $\hat{\boldsymbol{\beta}}_{(-i)}^T$ är oberoende stokastiska variabler, följer att även ε_i och $(\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})^T \boldsymbol{x}_i$ är oberoende. Från ekvationerna (8)-(9) får vi därför att

$$\begin{aligned} E(\text{PE}_i^2) &= E^2(\text{PE}_i) + \text{Var}(\text{PE}_i) \\ &= 0^2 + \text{Var}(\varepsilon_i - (\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})^T \boldsymbol{x}_i) \\ &= \text{Var}(\varepsilon_i) + \text{Var}[(\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta})^T \boldsymbol{x}_i] \\ &= \sigma^2 + \text{Var}(\textstyle\sum_{j=1}^m (\hat{\beta}_{(-i),j} - \beta_j) x_{ij}) \\ &= \sigma^2 + \text{Var}(\textstyle\sum_{j=1}^m \hat{\beta}_{(-i),j} x_{ij}) \\ &= \sigma^2 + \textstyle\sum_{j,k=1}^m x_{ij} x_{ik} \text{Cov}(\hat{\beta}_{(-i),j}, \hat{\beta}_{(-i),k}) \\ &= \sigma^2 + \boldsymbol{x}_i^T \boldsymbol{C}_i \boldsymbol{x}_i, \end{aligned} \tag{10}$$

där $\hat{\boldsymbol{\beta}}_{(-i)} = (\hat{\beta}_{(-i),1}, \ldots, \hat{\beta}_{(-i),m})^T$. Genom att sätta in (8) i uttrycket för CV, och sedan använda (10), får vi att

$$\begin{aligned} E(\text{CV}) &= \tfrac{1}{N} \textstyle\sum_{i=1}^N E(\text{PE}_i^2) \\ &= \tfrac{1}{N} \textstyle\sum_{i=1}^N (\sigma^2 + \boldsymbol{x}_i^T \boldsymbol{C}_i \boldsymbol{x}_i) \\ &= \sigma^2 + \tfrac{1}{N} \textstyle\sum_{i=1}^N \boldsymbol{x}_i^T \boldsymbol{C}_i \boldsymbol{x}_i. \end{aligned} \tag{11}$$

c) Designmatrisen är en $N \times m$-matris som ges av

$$\boldsymbol{X} = (x_{ji};\, i = 1, \ldots, N, j = 1, \ldots, m) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T.$$

Vidare gäller att

$$\boldsymbol{C} = \text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}. \tag{12}$$

d) Genom att använda ledningen, så ersätter vi $\boldsymbol{C}_i = \text{Var}(\hat{\boldsymbol{\beta}}_{(-i)})$ i (11) med uttrycket för $\boldsymbol{C}$ i (12) . Det ger

$$\begin{aligned} E(\text{CV}) &\approx \sigma^2 + \tfrac{1}{N} \textstyle\sum_{i=1}^N \boldsymbol{x}_i^T \boldsymbol{C} \boldsymbol{x}_i \\ &= \sigma^2 [1 + \tfrac{1}{N} \textstyle\sum_{i=1}^N \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i] \\ &= \sigma^2 (1 + \tfrac{1}{N} \textstyle\sum_{i=1}^N h_{ii}) \\ &= \sigma^2 (1 + \tfrac{m}{N}), \end{aligned} \tag{13}$$

där $\boldsymbol{H} = (h_{ij})_{i,j=1}^{N} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ är hattmatrisen. I sista ledet använde vi ledningen, att summan av diagonalelementen i hattmatrisen är $\sum_{i=1}^{N} h_{ii} = m$.

e) Om $\boldsymbol{\beta}$ är känd så kan vi prediktera Y_i med $\boldsymbol{\beta}^T\boldsymbol{x}_i$, så att prediktionsfelet för observation i är lika med dess felterm

$$\mathrm{PE}_i = Y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i = \varepsilon_i.$$

Det medför att

$$E(\mathrm{CV}) = \frac{1}{N}\sum_{i=1}^{N} E(\varepsilon_i^2) = \frac{1}{N}\sum_{i=1}^{N} \sigma^2 = \sigma^2$$

då $\boldsymbol{\beta}$ är känd. Således kan vi tolka $m\sigma^2/N$ som en bestraffningsterm som anger hur mycket $E(\mathrm{CV})$ ökar för att vi måste skatta m parametrar i $\boldsymbol{\beta}$.