

**Lösningar till tentamensskrivning för kursen
Introduktion till övervakad statistisk inläring**

25 november 2025 8–13

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

Uppgift 1

a) Skriv den centrerade regressionsmodellen som

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \dots, 16,$$

med intercept $\tilde{\alpha} = \alpha + \beta\bar{x}$. Minsta kvadrat-skattningarna av $\tilde{\alpha}$ och β ges av

$$\begin{aligned}\hat{\tilde{\alpha}} &= \sum_i Y_i / 16 = 312.0 / 16 = 19.5, \\ \hat{\beta} &= \sum_i Y_i (x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 = 9.0 / 45.0 = 0.20.\end{aligned}\tag{1}$$

Det ger en skattad halt

$$\hat{\alpha} = \hat{\tilde{\alpha}} - \hat{\beta} \cdot \bar{x} = 19.5 - 0.2 \cdot \frac{168.0}{16} = 17.4\tag{2}$$

av typ A-bakterier.

b) Eftersom de två skattningarna i (1) är oberoende stokastiska variabler, följer av (2) att

$$\begin{aligned}\text{Var}(\hat{\alpha}) &= \text{Var}(\hat{\tilde{\alpha}} - \hat{\beta} \cdot \bar{x}) \\ &= \text{Var}(\hat{\tilde{\alpha}}) + \text{Var}(\hat{\beta}) \cdot \bar{x}^2 \\ &= \frac{\sigma^2}{16} + \frac{\sigma^2 \bar{x}^2}{\sum_i (x_i - \bar{x})^2} \\ &= \sigma^2 \left(\frac{1}{16} + \frac{(168.0/16)^2}{45.0} \right) \\ &= 2.5125 \cdot \sigma^2.\end{aligned}\tag{3}$$

För att skatta feltermernas varians så utnyttjar vi att variationskällan Residual har $16-2=14$ frihetsgrader. Av detta följer att

$$\hat{\sigma}^2 = \frac{\text{Kvs(Residual)}}{14} = \frac{19.0}{14} = 1.3571.\tag{4}$$

Genom att kombinera (3) med (4) så får vi ett medelfel

$$d = \sqrt{2.5125} \cdot \hat{\sigma} = \sqrt{2.5125 \cdot 1.3571} = 1.8465 = 1.85.$$

c) Ett 95 % konfidensintervall för halten A-bakterier är

$$\begin{aligned} I_\alpha &= (\hat{\alpha} - t_{0.025}(14) \cdot d, \hat{\alpha} + t_{0.025}(14) \cdot d) \\ &= (17.4 - 2.145 \cdot 1.8465, 17.4 + 2.145 \cdot 1.8465) \\ &= (13.4, 21.4), \end{aligned}$$

där värdet på t -kvantilen fås från tabell ($t_{0.025}(14) = \sqrt{F_{0.05}(1, 14)}$).

Uppgift 2

a) Vi börjar med att fylla i antalet frihetsgrader f i den fullständiga modellens variansanalystabell, med förkortningarna M1, M2 och M3 för de tre metallerna:

Variationskälla	f	Kvs
M1	1	3.1
M2	1	5.1
M3	1	8.9
Residual	16	25.0
Totalt	19	42.1

b) I första FS-steget testas tre olika grundmodeller, var och en med en förklarande variabel, med ett F -test mot en hypotesmodell som bara innehåller intercept. Eftersom M3 har störst kvadratsumma börjar vi med att undersöka F -kvoten för delmodellen som endast har M3 som förklarande variabel. Om vi använder denna delmodell som grundmodell får vi enligt ledningen (med Regression = Avvikelse mellan grund- och hypotesmodell)

$$\begin{aligned} \text{Kvs(Regression)} &= \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}\|^2 \\ &= \text{Kvs(M3)} \\ &= 8.9, \\ \text{Kvs(Residual)} &= \text{Kvs(Residual)}_{\text{Modell med M3}} \\ &= \|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 \\ &= \text{Kvs(M1)} + \text{Kvs(M2)} + \text{Kvs(Residual)}_{\text{Fullst modell}} \\ &= 3.1 + 5.1 + 25.0 \\ &= 33.2, \end{aligned}$$

där $\mathbf{Y}$ är observationsvektorn, och $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\mu}}$ skattningar av dess väntevärde enligt grund- respektive hypotesmodellen. Eftersom antalet frihetsgrader för Regression och Residual är 1 respektive $1 + 1 + 16 = 18$ får vi en

$$\begin{aligned} F\text{-kvot} &= \frac{\text{Kvs(Regression)}/1}{\text{Kvs(Residual)}/18} \\ &= \frac{8.9}{33.2/18} \\ &= 4.825, \end{aligned}$$

som överstiger $F_{0.05}(1, 18) = 4.41$. Därför förkastas nollhypotesen, svarande mot att M3 ger ett signifikant bidrag till legeringens hållfasthet.

De andra två delmodellerna med bara M1 respektive M2 som förklarande variabler har också 1 frihetsgrad för Regression och 18 frihetsgrader för Residual. Eftersom deras kvadratsummor $Kvs(\text{Regression}) = Kvs(\text{M1})$ respektive $Kvs(\text{Regression}) = Kvs(\text{M2})$ för regression är lägre och deras kvadratsummor $Kvs(\text{Residual}) = Kvs(\text{Total}) - Kvs(\text{Regression})$ för residual är högre jämfört med modellen som bara har M3 som förklarande variabel, så kommer F -kvoterna för båda dessa delmodeller ha lägre värden än 4.825. Därför väljer vi M3 i första steget av FS-schemat.

c) Eftersom vi tog med M3 som förklarande variabel i 2b) så går vi vidare i andra steget av FS-schemat och testar hypotesmodellen med bara M3 som förklarande variabel mot två olika grundmodeller, som även inkluderar M1 respektive M2. Eftersom M2 har större kvadratsumma än M1, och alla prediktorerna är ortogonala, räcker det att undersöka om M2 ska tas med utöver M3, av samma skäl som att det i 2b) räckte att testa grundmodellen med M3 mot hypotesmodellen med endast intercept. Med M2 och M3 i grundmodellen och bara M3 i hypotesmodellen får vi då på grund av ortogonaliteten mellan prediktorerna (se ledningen)

$$\begin{aligned}
 \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 &= Kvs(\text{Regression})_{\text{Modell med M2 och M3}} \\
 &- Kvs(\text{Regression})_{\text{Modell med M3}} \\
 &= [Kvs(\text{M2}) + Kvs(\text{M3})] - Kvs(\text{M3}) \\
 &= Kvs(\text{M2}) \\
 &= 5.1, \\
 Kvs(\text{Residual}) &= Kvs(\text{Residual})_{\text{Modell med M2 och M3}} \\
 &= \|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 \\
 &= Kvs(\text{M1}) + Kvs(\text{Residual})_{\text{Fullst modell}} \\
 &= 3.1 + 25.0 \\
 &= 28.1,
 \end{aligned} \tag{5}$$

och en

$$\begin{aligned}
 F\text{-kvot} &= \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/1}{Kvs(\text{Residual})/(1+16)} \\
 &= \frac{5.1}{28.1/17} \\
 &= 3.085,
 \end{aligned} \tag{6}$$

som understiger $F_{0.05}(1, 17) = 4.451$. Därför tas inte M2 med i modellen. Slutsatsten blir att FS-schemat inte tar med någon fler förklarande variabel i andra steget utan stannar efter det första steget. Den valda modellen har alltså bara med M3 som förklarande variabel.

Uppgift 3

a) Antalet frihetsgrader för residualerna $Y_{ij} - \bar{Y}_{i\cdot}$ är $6(4 - 1) = 18$. Det ger en väntevärdesriktig skattning

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 &= \text{Mvs(Inom blodkropp)} \\ &= \frac{1}{18} \text{Kvs(Inom blodkropp)} \\ &= \frac{1}{18} \sum_{i=1}^6 \sum_{j=1}^4 (Y_{ij} - \bar{Y}_{i\cdot})^2 \\ &= \frac{1}{18} \sum_{i=1}^6 3s_i^2 \\ &= \frac{1}{6} \sum_{i=1}^6 s_i^2 \\ &= \frac{1}{6} (0.020 + 0.015 + 0.025 + 0.018 + 0.022 + 0.014) \\ &= 0.019\end{aligned}$$

av σ_ε^2 . Eftersom radmedelvärdena

$$\bar{Y}_{i\cdot} = \mu + \delta_i + \bar{\varepsilon}_{i\cdot} \sim N(\mu, \sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{4})$$

är sinsemellan oberoende, kan vi använda deras stickprovsvarians för att beräkna en väntevärdesriktig skattning

$$\begin{aligned}\hat{\sigma}_\delta^2 + \frac{1}{4}\hat{\sigma}_\varepsilon^2 &= \frac{1}{6-1} \sum_{i=1}^6 (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= \frac{1}{5} [(7.5 - 7.6)^2 + (7.1 - 7.6)^2 + (8.2 - 7.6)^2 + (7.4 - 7.6)^2 \\ &\quad + (7.9 - 7.6)^2 + (7.5 - 7.6)^2] \\ &= 0.152\end{aligned}$$

av $\sigma_\delta^2 + \sigma_\varepsilon^2/4$. Det ger i sin tur en väntevärdesriktig skattning

$$\hat{\sigma}_\delta^2 = 0.152 - \frac{0.019}{4} = 0.147$$

av σ_δ^2 .

b) Minsta kvadrat-skattningen av μ är $\hat{\mu} = \bar{Y}_{\cdot\cdot} = 7.6$. Eftersom

$$\text{Var}(\hat{\mu}) = \frac{\sigma_\delta^2}{6} + \frac{\sigma_\varepsilon^2}{24}$$

kan vi utnyttja räkningarna i deluppgift 3a) för att erhålla ett medelfel

$$\begin{aligned}d &= \sqrt{\widehat{\text{Var}}(\hat{\mu})} \\ &= \sqrt{\frac{1}{6}(\hat{\sigma}_\delta^2 + \frac{1}{4}\hat{\sigma}_\varepsilon^2)} \\ &= \sqrt{\frac{0.152}{6}} \\ &= 0.1592,\end{aligned}$$

och ett konfidensintervall

$$\begin{aligned}I_\mu &= \hat{\mu} \pm t_{0.025}(5)d \\ &= 7.6 \pm \sqrt{F_{0.05}(1, 5)} \cdot 0.1592 \\ &= 7.6 \pm \sqrt{6.608} \cdot 0.1592 \\ &= (7.19, 8.01)\end{aligned}$$

för μ med konfidensgrad 95%.

Uppgift 4

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

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

Genom att lösa (7) med avseende på γ_1 får vi

$$\gamma_1 = \phi_1 \gamma_0 / (1 - \phi_2) \iff \rho_1 = \gamma_1 / \gamma_0 = \phi_1 / (1 - \phi_2).\tag{8}$$

Vi fortsätter med $k = 2$ och noterar att

$$\begin{aligned}\gamma_2 &= \text{Cov}(X_t, X_{t+2}) = \text{Cov}(X_t, \phi_1 X_{t+1} + \phi_2 X_t + \varepsilon_{t+2}) \\ &= \phi_1 \gamma_1 + \phi_2 \gamma_0 = \gamma_0 (\phi_1 \rho_1 + \phi_2) \\ &= \gamma_0 (\phi_1^2 / (1 - \phi_2) + \phi_2),\end{aligned}\tag{9}$$

där vi i sista ledet utnyttjade (8). Division med γ_0 i båda leden av (9) leder slutligen till

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

b) Låt $\bar{y} = \hat{\mu} = \sum_{t=1}^T y_t / T$ vara stickprovsmedelvärdet av den observerade tidsserien. Skattningen av kovariansfunktionen ges av

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (y_t - \bar{y})(y_{t+k} - \bar{y}),$$

vilket i sin tur ger en skattning

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{t=1}^{T-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$$

av autokorrelationsfunktionen (korrelogrammet).

c) Eftersom $Y_t = \mu + \varepsilon_t$ är oberoende under nollhypotesen så gäller approximativt, för stora T , att $\hat{\rho}_1$ och $\hat{\rho}_2$ är oberoende med $\sqrt{T} \hat{\rho}_k \sim N(0, 1)$. Det medför att vi approximativt har

$$Q = T(\hat{\rho}_1^2 + \hat{\rho}_2^2) \sim \chi^2(2)$$

under H_0 . Ett hypotestest som förkastar H_0 då $Q > \chi_{0.05}^2(2) = 5.99$ har alltså approximativt signifikansnivån 5%.

Uppgift 5

a) Vi skriver om den allmänna linjära modellen för datamaterialet på matrisform $\mathbf{Y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, där $\mathbf{Y} = (Y_1, \dots, Y_6)^T$ är observationsvektorn, $\boldsymbol{\theta} = (\beta_1, \beta_2)^T$ parametervektorn, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_6)^T$ feltermsvektorn och

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ -1 & 1 \\ -1 & -1 \\ 1 & -1 \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2).$$

designmatrisen. Speciellt ser vi att designmatrisens två kolumner $\mathbf{x}_1$ och $\mathbf{x}_2$ är ortogonala, så att

$$\mathbf{A}^T \mathbf{A} = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$$

är diagonal. Minsta kvadrat-skattaren av effektparametrarna kan skrivas som

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Y} = \frac{1}{4} \mathbf{A}^T \mathbf{Y} = \begin{pmatrix} \mathbf{x}_1^T \mathbf{Y} / 4 \\ \mathbf{x}_2^T \mathbf{Y} / 4 \end{pmatrix}. \quad (10)$$

Vidare ges den simultana normalfördelningen för $\hat{\beta}_1$ och $\hat{\beta}_2$ av

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 (\mathbf{A}^T \mathbf{A})^{-1} \right) = N \left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \sigma^2/4 & 0 \\ 0 & \sigma^2/4 \end{pmatrix} \right). \quad (11)$$

b) Vi utnyttjar ledningen och skriver om prediktionsfelet som

$$Y - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 = \varepsilon - (\hat{\beta}_1 - \beta_1)x_1 - (\hat{\beta}_2 - \beta_2)x_2.$$

Enligt 5a) är $\hat{\beta}_1$ och $\hat{\beta}_2$ sinsemellan oberoende, eftersom kovariansmatrisen i (11) är diagonal. Eftersom $\hat{\beta}_1$ och $\hat{\beta}_2$ endast beror av feltermerna $\varepsilon_1, \dots, \varepsilon_6$, är dessa skattningar oberoende av feltermen ε för den nya observationen. Vi får därför att

$$\begin{aligned} \text{MSEP} &= \text{Var}(\varepsilon) + x_1^2 \text{Var}(\hat{\beta}_1) + x_2^2 \text{Var}(\hat{\beta}_2) \\ &= \sigma^2 \left[1 + \frac{x_1^2 + x_2^2}{4} \right]. \end{aligned} \quad (12)$$

c) Den förenklade modellen, där bara kovariat 1 ingår, kan skrivas som $\mathbf{Y} = \mathbf{A}_1 \beta_1 + \boldsymbol{\epsilon}$, där designmatrisen $\mathbf{A}_1 = \mathbf{x}_1$ svarar mot första kolumnen i $\mathbf{A}$ och där $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_6)^T$. Vi får därför en minsta kvadrat-skattning

$$(\mathbf{A}_1^T \mathbf{A}_1)^{-1} \mathbf{A}_1^T \mathbf{Y} = \frac{1}{4} \mathbf{x}_1^T \mathbf{Y} = \hat{\beta}_1$$

av β_1 som överensstämmer med (10), eftersom kolumnerna i $\mathbf{A}$ är ortogonala. Prediktionsfelet för den nya observationen, när det första datasetet anpassats till den mindre modellen med kovariat 1, blir därför

$$Y - \hat{\beta}_1 x_1 = \varepsilon - (\hat{\beta}_1 - \beta_1)x_1 + \beta_2 x_2.$$

På motsvarande sätt som i 5a) får vi ett prediktionsfel

$$\begin{aligned} \text{MSEP}_1 &= \text{Var}(\varepsilon) + x_1^2 \text{Var}(\hat{\beta}_1) + x_2^2 \beta_2^2 \\ &= \sigma^2 \left[1 + \frac{x_1^2}{4} \right] + x_2^2 \beta_2^2. \end{aligned} \tag{13}$$

Genom att bilda differensen mellan (12) och (13) ser vi att

$$\text{MSEP} - \text{MSEP}_1 = x_2^2 \left(\frac{\sigma^2}{4} - \beta_2^2 \right).$$

Alltså ger den mindre modellen en bättre prediktion av Y om $x_2 \neq 0$ och $\sigma^2 > 4\beta_2^2$. Det lönar sig alltså inte att skatta β_2 om denna parameter är liten i förhållande till feltermsvariansen σ^2 .