

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

24 oktober 2025 8–13

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

Uppgift 1

a) Låt $j = 1, 2, 3, 4, 5$ beteckna gruppnummer, n_j antal personer och $\bar{Y}_j$ medelvärdet av Y_i för alla personer i grupp j . Eftersom interceptet $\tilde{\alpha}$ är centrerat följer att

$$\begin{aligned}\hat{\tilde{\alpha}} &= \bar{Y} = \sum_{i=1}^{30} Y_i / 30 \\ &= (n_1 \bar{Y}_1 + n_2 \bar{Y}_2 + n_3 \bar{Y}_3 + n_4 \bar{Y}_4 + n_5 \bar{Y}_5) / 30 \\ &= (4 \cdot 19.0 + 6 \cdot 22.0 + 10 \cdot 24.0 + 6 \cdot 25.5 + 4 \cdot 26.5) / 30 \\ &= 23.57.\end{aligned}$$

Med hjälp av ledningen får vi direkt att skattningen av lutningsparametern β ges av

$$\begin{aligned}\hat{\beta} &= \sum_{i=1}^{30} (x_i - \bar{x}) Y_i / \sum_{i=1}^{30} (x_i - \bar{x})^2 \\ &= 81/44 = 1.841,\end{aligned}$$

där

$$\begin{aligned}\bar{x} &= \sum_{i=1}^{30} x_i / 30 \\ &= (n_1 \cdot 5 + n_2 \cdot 6 + n_3 \cdot 7 + n_4 \cdot 8 + n_5 \cdot 9) / 30 \\ &= (4 \cdot 5 + 6 \cdot 6 + 10 \cdot 7 + 6 \cdot 8 + 4 \cdot 9) / 30 \\ &= 7.\end{aligned}$$

b) Låt $N = 30$ vara antalet observationer. Skattningsvektorn $(\hat{\tilde{\alpha}}, \hat{\beta})^T$ är tvådimensionellt normalfördelad med väntevärde $(\tilde{\alpha}, \beta)^T$ och kovariansmatris

$$\begin{pmatrix} \sigma^2/N & 0 \\ 0 & \sigma^2 / \sum_{i=1}^N (x_i - \bar{x})^2 \end{pmatrix} = \begin{pmatrix} \sigma^2/30 & 0 \\ 0 & \sigma^2/44 \end{pmatrix}.$$

c) Eftersom den enkla linjära regressionsmodellen innehåller 2 parametrar är antalet frihetsgrader för variationskällan Residual lika med $30 - 2 = 28$. Det ger en väntevärdesriktig skattning

$$\hat{\sigma}^2 = \text{Mkvs(Residual)} = \frac{\text{Kvs(Residual)}}{28} = \frac{550}{28} = 19.64$$

av variansparametern σ^2 .

d) En person som sovit 6.5 timmar har en förväntad minnesförmåga

$$\mu = \tilde{\alpha} + (6.5 - \bar{x})\beta = \alpha - 0.5\beta.$$

Motsvarande skattning

$$\hat{\mu} = \hat{\tilde{\alpha}} - 0.5\hat{\beta} = 23.57 - 0.5 \cdot 1.841 = 22.65$$

har variansen

$$\text{Var}(\hat{\mu}) = \text{Var}(\hat{\tilde{\alpha}}) + 0.25\text{Var}(\hat{\beta}) = \frac{\sigma^2}{30} + \frac{0.25\sigma^2}{44} = 0.039\sigma^2.$$

Det ger ett medelfel

$$d = \sqrt{\widehat{\text{Var}}(\hat{\mu})} = \sqrt{0.039\hat{\sigma}^2} = \sqrt{0.039 \cdot 19.64} = 0.8753$$

för $\hat{\mu}$ och ett 95% konfidensintervall

$$\begin{aligned} (\hat{\mu} - t_{0.025}(28)d, \hat{\mu} + t_{0.025}(28)d) &= (22.65 - 2.0484 \cdot 0.8753, 22.65 + 2.0484 \cdot 0.8753) \\ &= (20.85, 24.44) \end{aligned}$$

för μ . Här fås t -kvantilen ur tabell med F -fördelningens kvantiler genom $t_{0.025}(28) = \sqrt{F_{0.05}(1, 28)}$.

Uppgift 2

a) Modellen kan skrivas som

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad (1)$$

för syreupptagningsförmågan hos person $k \in \{1, \dots, 4\}$ inom gruppen för vilka rökning är på nivån $i \in \{1, 2, 3\}$ och den fysiska aktiviteten på nivån $j \in \{1, 2\}$. Vidare är μ det genomsnittliga väntevärdet för alla grupper, α_i den systematiska effekten av rökning på nivå i , β_j den systematiska effekten av fysisk aktivitet på nivå j samt γ_{ij} samspelet mellan rökning och fysisk aktivitet. För att undvika överparametrisering inför vi totalt 6 linjärt oberoende bivillkor $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ (det vill säga 1 bivillkor för α_i , 1 för β_j och $3+2-1=4$ för γ_{ij}). Feltermerna $\varepsilon_i \sim N(0, \sigma^2)$ antas vara oberoende.

b) För att testa grundmodellen (1) mot hypotesmodellen

$$\gamma_{ij} = 0, \forall i, j$$

att det inte finns något samspel mellan rökning och fysisk aktivitet, bildar vi

$$\text{F-kvot} = \frac{\text{Mkvs}(\text{Samspel})}{\text{Mkvs}(\text{Inom celler})} = \frac{\text{Kvs}(\text{Samspel})/2}{\text{Kvs}(\text{Inom celler})/18} = \frac{5.5/2}{19.5/18} = 2.54.$$

Här utnyttjade vi att variationskällan Samspel har $(2 - 1)(3 - 1) = 2$ frihetsgrader, medan Inom celler har $3 \cdot 2(4 - 1) = 18$ frihetsgrader. Då F-kvoten har en $F(2, 18)$ -fördelning under H_0 så jämför vi dess observerade värde med

$$F_{0.05}(2, 18) = 3.55.$$

Eftersom F-kvoten inte överstiger detta värde kan vi inte förkasta H_0 på signifikansnivån 5%.

c) Eftersom samsplet i 2b) inte var signifikant så antar vi en additiv modell (=hypotesmodellen i 2b). Alltså slår vi ihop de två variationskällorna Samspel och Inom celler till en ny variationskälla med $2+18=20$ frihetsgrader. Vi skattar sedan feltermernas varians enligt

$$\hat{\sigma}^2 = \frac{\text{Kvs}(\text{Samspel}) + \text{Kvs}(\text{Inom celler})}{2 + 18} = \frac{5.5 + 19.5}{20} = 1.25.$$

Eftersom variationskällan Rökning har $3-1=2$ frihetsgrader får vi en

$$\text{F-kvot} = \frac{\text{Kvs}(\text{Rökning})/2}{\hat{\sigma}^2} = \frac{10.0/2}{1.25} = 4.0 > F_{0.05}(2, 20) = 3.48.$$

Således kan vi förkasta nollhypotesen att rökning inte har någon effekt på syreupptagningsförmågan, på nivån 5%.

Uppgift 3

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

b) Sätt $X_t = Y_t - \mu$, så att

$$X_t = \phi X_{t-1} + \varepsilon_t. \quad (2)$$

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

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

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

$$\begin{aligned} \gamma_k &= \text{Cov}(Y_t, Y_{t+k}) = \text{Cov}(X_t, X_{t+k}) = \text{Cov}(X_t, \phi X_{t+k-1} + \varepsilon_{t+k}) \\ &= \phi \text{Cov}(X_t, X_{t+k-1}) + \text{Cov}(X_t, \varepsilon_{t+k}) = \phi \gamma_{k-1}. \end{aligned} \quad (3)$$

Eftersom $\gamma_0 = \sigma^2$ så medför (3) (med induktionsbevis) att $\gamma_k = \gamma_0 \phi^k$. Det ger

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

c) Eftersom

$$Y_{T+1} = \mu + \phi(Y_T - \mu) + \varepsilon_{T+1}$$

så kan vi utnyttja ledningen för att dra slutsatsen att

$$\begin{aligned}\hat{Y}_{T+1} &= E(Y_{T+1}|Y_T) \\ &= E(\mu + \phi(Y_T - \mu) + \varepsilon_{T+1}|Y_T) \\ &= \mu + \phi(Y_T - \mu) + E(\varepsilon_{T+1}|Y_T) \\ &= \mu + \phi(Y_T - \mu),\end{aligned}$$

där vi i sista ledet utnyttjade att Y_T och ε_{T+1} är oberoende (se ledningen till uppgift 3b), och alltså $E(\varepsilon_{T+1}|Y_T) = E(\varepsilon_{T+1}) = 0$.

Uppgift 4

a) Den givna modellen (ekvation (3) i skrivningsbladet) kan skrivas på matrisform som $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, där

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} Z_1 - 0.5 \\ Z_2 - 0.5 \\ Z_3 - 0.5 \\ Z_4 - 0.5 \end{pmatrix} = \begin{pmatrix} -0.27 \\ -0.09 \\ 0.12 \\ 0.24 \end{pmatrix},$$

är observationsvektorn,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \\ x_{14} & x_{24} \end{pmatrix} = \begin{pmatrix} -0.5 & -0.5 \\ 0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

är designmatrisen och $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)^T$ feltermsvektorn. Vi börjar med att räkna ut

$$\mathbf{S} = \mathbf{X}^T \mathbf{X} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Det ger en minsta-kvadratskattning

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \mathbf{S}^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{Y} = \begin{pmatrix} 0.5(-Y_1 + Y_2 - Y_3 + Y_4) \\ 0.5(-Y_1 - Y_2 + Y_3 + Y_4) \end{pmatrix} = \begin{pmatrix} 0.15 \\ 0.36 \end{pmatrix}.$$

b) Kovariansmatrisen för $\hat{\boldsymbol{\beta}}$ ges av

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{S}^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}.$$

Variationsinflationsfaktorn för $\hat{\beta}_1$ anger hur mycket variansen av skattningen av β_1 ökar på grund av att man även måste skatta β_2 . Eftersom variansen för skattningen av β_1 är σ^2/s_{11} då β_2 är känd, och $\sigma^2(\mathbf{S}^{-1})_{11}$ då β_2 är okänd, följer att

$$\text{VIF}(\hat{\beta}_1) = \frac{\sigma^2(\mathbf{S}^{-1})_{11}}{\sigma^2/s_{11}} = s_{11} \cdot (\mathbf{S}^{-1})_{11} = 1 \cdot 1 = 1.$$

Variationsinflationsfaktorn är alltså 1 eftersom de två förklarande variablerna x_1 och x_2 är ortogonala.

c) Vi börjar med att skatta feltermsvariansen. Eftersom residualerna har $N - 2 = 4 - 2 = 2$ frihetsgrader följer av ledningen och de uträknade skattningarna av β_1 och β_2 i 4a), att

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{2} \sum_{i=1}^4 (Y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 \\ &= \frac{1}{2} \left(\sum_{i=1}^4 Y_i^2 - \hat{\beta}_1^2 \sum_{i=1}^4 x_{1i}^2 - \hat{\beta}_2^2 \sum_{i=1}^4 x_{2i}^2 \right) \\ &= \frac{1}{2} \left(\sum_{i=1}^4 Y_i^2 - \hat{\beta}_1^2 s_{11} - \hat{\beta}_2^2 s_{22} \right) \\ &= \frac{1}{2} (0.153 - 0.15^2 - 0.36^2) \\ &= 4.5 \cdot 10^{-4}. \end{aligned}$$

Låt $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ vara väntevärdesvektorn för observationerna. Varje värde på $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ kan testas som en nollhypotes, baserat på en

$$\text{F-kvot} = \frac{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 / (2 - 0)}{\hat{\sigma}^2} = \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2}{2\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{S}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{2\hat{\sigma}^2}.$$

Eftersom $\mathbf{S} = \mathbf{X}^T \mathbf{X} = \mathbf{I}_2$ är enhetsmatrisen av ordning 2, enligt 4a), så följer att

$$\text{F-kvot} = \frac{(\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2}{2 \cdot 4.5 \cdot 10^{-4}}.$$

Denna F-kvot har en $F(2, 2)$ -fördelning under nollhypotesen. Det ger en konfidsensregion E med konfidsensgrad 95% som består av alla värden på $(\beta_1, \beta_2)^T$ för vilka nollhypotesen inte förkastas, det vill säga de värden på $(\beta_1, \beta_2)^T$ för vilka F-kvoten ovan inte överstiger $F_{0.05}(2, 2) = 19.0$. Det ger

$$\begin{aligned} E &= \{(\beta_1, \beta_2)^T; (\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2 \leq 2 \cdot 4.5 \cdot 10^{-4} \cdot 19.0\} \\ &= \{(\beta_1, \beta_2)^T; (0.15 - \beta_1)^2 + (0.36 - \beta_2)^2 \leq 0.0171\}. \end{aligned}$$

Uppgift 5

a) Förklaringsgraderna för grund- respektive hypotesmodellerna anger hur stor andel av variationen i responsvariablerna Y_i som fångas upp av $\hat{\mu}_i$ respektive $\hat{\hat{\mu}}_i$. Det svarar mot $\text{Kvs}(\text{Regression})/\text{Kvs}(\text{Total})$ för respektive modell, dvs

$$R_0^2 = \frac{\sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2} = \frac{\|\hat{\boldsymbol{\mu}} - \bar{Y}\|^2}{\|\mathbf{Y} - \bar{Y}\|^2} \quad (4)$$

och

$$R_1^2 = \frac{\sum_{i=1}^N (\hat{\mu}_i - \bar{Y})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2} = \frac{\|\hat{\boldsymbol{\mu}} - \bar{\mathbf{Y}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2}. \quad (5)$$

Vi införde här observationsvektorn $\mathbf{Y} = (Y_1, \dots, Y_N)^T$ och vektorn $\bar{\mathbf{Y}} = (\bar{Y}, \dots, \bar{Y})^T$ som har identiska koordinater lika med skattningen av interceptet (dvs $\hat{\alpha} = \bar{Y}$).

b) Vektorerna $\hat{\boldsymbol{\mu}}$ och $\hat{\hat{\boldsymbol{\mu}}}$ är projektioner av observationsvektorn $\mathbf{Y}$ ned på de delrum $\mathcal{U}_k$ och $\mathcal{U}_l \subset \mathcal{U}_k$ av dimension $k = m + 1$ och $l = m$ som svarar mot grund- respektive hypotesmodellerna. Därför kommer även $\hat{\hat{\boldsymbol{\mu}}}$ vara projektionen av $\hat{\boldsymbol{\mu}}$ ned på hypotesrummet $\mathcal{U}_l$. Eftersom hypotesmodellen innehåller intercept så gäller $\bar{\mathbf{Y}} \in \mathcal{U}_l$ och därmed också $\hat{\boldsymbol{\mu}} - \bar{\mathbf{Y}} \in \mathcal{U}_l$. Eftersom $\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}$ är ortogonal mot alla element i $\mathcal{U}_l$ så är $\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}$ ortogonal mot $\hat{\hat{\boldsymbol{\mu}}} - \bar{\mathbf{Y}}$. Av detta följer att

$$\|\hat{\boldsymbol{\mu}} - \bar{\mathbf{Y}}\|^2 = \|(\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}) + (\hat{\hat{\boldsymbol{\mu}}} - \bar{\mathbf{Y}})\|^2 = \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 + \|\hat{\hat{\boldsymbol{\mu}}} - \bar{\mathbf{Y}}\|^2.$$

Genom insättning i (4)-(5) ger det i sin tur att

$$R_0^2 - R_1^2 = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2} = \frac{\sum_{i=1}^N (\hat{\mu}_i - \hat{\hat{\mu}}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}. \quad (6)$$

c) Eftersom minsta kvadrat-skattningen av $\boldsymbol{\theta}$ ges av $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_m)^T = (\bar{Y}, \hat{\beta}_1, \dots, \hat{\beta}_m)^T$, där $\hat{\beta}_j$ svarar mot kolumnen $\mathbf{x}_j - \bar{\mathbf{x}}_j$ i designmatrisen $\mathbf{A}$, så följer att

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\hat{\boldsymbol{\theta}} = \bar{\mathbf{Y}} + \sum_{l=1}^m \hat{\beta}_l(\mathbf{x}_l - \bar{\mathbf{x}}_l) = \hat{\beta}_j \mathbf{x}_j + \mathbf{v}, \quad (7)$$

där $\mathbf{v} \in \mathcal{U}_l$. Det beror på att $\bar{\mathbf{x}}_j$ och alla kolumner i $\mathbf{A}$ som inte svarar mot kovariat j , tillhör hypotesrummet $\mathcal{U}_l$. Nu är $\hat{\hat{\mathbf{x}}}_j$ en projektion av $\mathbf{x}_j$ ned på hypotesrummet $\mathcal{U}_l$. Således är $\mathbf{x}_j - \hat{\hat{\mathbf{x}}}_j$, och därmed även $\hat{\beta}_j(\mathbf{x}_j - \hat{\hat{\mathbf{x}}}_j)$, ortogonal mot $\mathcal{U}_l$. Vi skriver om (7) som

$$\hat{\boldsymbol{\mu}} = \hat{\beta}_j(\mathbf{x}_j - \hat{\hat{\mathbf{x}}}_j) + \mathbf{w}, \quad (8)$$

där $\mathbf{w} = \mathbf{v} + \hat{\beta}_j \hat{\hat{\mathbf{x}}}_j \in \mathcal{U}_l$, eftersom $\mathbf{v} \in \mathcal{U}_l$ och $\hat{\hat{\mathbf{x}}}_j \in \mathcal{U}_l$. Men eftersom projektionen av $\hat{\boldsymbol{\mu}}$ ned på hypotesrummet är $\hat{\hat{\boldsymbol{\mu}}}$, och $\hat{\beta}_j(\mathbf{x}_j - \hat{\hat{\mathbf{x}}}_j)$ är ortogonal mot $\mathcal{U}_l$, så följer att $\mathbf{w} = \hat{\hat{\boldsymbol{\mu}}}$ i (8). Därmed har vi visat att

$$\hat{\mu}_i = \hat{\beta}_j(x_{ji} - \hat{\hat{x}}_{ji}) + \hat{\hat{\mu}}_i$$

för $i = 1, \dots, N$. Insättning i (6) ger

$$R_0^2 - R_1^2 = \frac{\hat{\beta}_j^2 \sum_{i=1}^N (x_{ji} - \hat{\hat{x}}_{ji})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}.$$

Hur mycket mer vi förklarar med hjälp av $\mathbf{x}_j$ beror alltså dels på hur stor skattad effekt $\hat{\beta}_j$ denna variabel har och dels på hur stor del av kovariatvektorn $\mathbf{x}_j$ som inte förklaras av de övriga kovariaterna $\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_m$.