STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

17 april 2024 8–13

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

Uppgift 1

a) Skriv den centrerade regressionsmodellen som

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 12,$$

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

$$\begin{array}{rcl} \hat{\tilde{\alpha}} & = & \sum_i Y_i/12 = 31.4/12 = 2.6167, \\ \hat{\beta} & = & \sum_i Y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 = -4.1/13.2 = -0.3106. \end{array} \tag{1}$$

Det ger en skattad hållfasthet

$$\hat{\alpha} = \hat{\tilde{\alpha}} - \hat{\beta}\bar{x} = 2.6167 - (-0.3106) \cdot \frac{20.5}{12} = 3.147 \tag{2}$$

för legeringen.

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

$$\begin{array}{rcl} \mathrm{Var}(\hat{\alpha}) & = & \mathrm{Var}(\hat{\tilde{\alpha}}) + \mathrm{Var}(\hat{\beta}) \cdot \bar{x}^2 \\ & = & \frac{\sigma^2}{12} + \frac{\sigma^2 \bar{x}^2}{\sum_i (x_i - \bar{x})^2} \\ & = & \sigma^2 \left(\frac{1}{12} + \frac{(20.5/12)^2}{13.2} \right) \\ & = & 0.3044 \cdot \sigma^2. \end{array} \tag{3}$$

För att skatta feltermernas varians så börjar vi med att räkna ut kvadratsumman för variationskällan residual i en enkel linjär regressionsmodell. Vi får att

$$\begin{array}{rcl} \mathrm{Kvs(Residual)} & = & \mathrm{Kvs(Total)} - \mathrm{Kvs(Regression)} \\ & = & \sum_i (Y_i - \bar{Y})^2 - \hat{\beta}^2 \sum_i (x_i - \bar{x})^2 \\ & = & 3.3 - (-0.3106)^2 \cdot 13.2 \\ & = & 2.027. \end{array}$$

Eftersom antal frihetsgrader för Residual är 12-2=10, så följer att

$$\hat{\sigma}^2 = \frac{\text{Kvs(Residual)}}{10} = \frac{2.027}{10} = 0.2027. \tag{4}$$

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

$$d = \sqrt{0.3044} \cdot \hat{\sigma} = \sqrt{0.3044 \cdot 0.2027} = 0.2484.$$

c) Ett 95 % konfidensintervall för den okontaminerade metallegeringens hållfasthet är

$$\begin{aligned} I_\alpha &= (\hat{\alpha} - t_{0.025}(10) \cdot d, \hat{\alpha} + t_{0.025}(10) \cdot d) \\ &= (3.147 - 2.228 \cdot 0.2484, 3.147 + 2.228 \cdot 0.2484) \\ &= (2.59, 3.70), \end{aligned}$$

där värdet på t-kvantilen fås från tabell ($t_{0.025}(10) = \sqrt{F_{0.05}(1,10)} = \sqrt{5.0}$).

Uppgift 2

a) Antalet observationer är $N = 25$, hypotesmodellen har $l = 4$ parametrar (de tre livsstilsfaktorerna plus intercept) och grundmodellen har $k = 6$ parametrar. Att testa om de genetiska faktorerna har någon signifikant inverkan på åldersdiabetes (utöver livsstilsfaktorerna) är samma sak som att testa avvikelse från hypotesmodellen. För detta ändamål används

$$\begin{aligned} \text{F-kvot} &= \frac{\text{Mkvs(Avv från hypotes)}}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Avv från hypotes)}/(k-l)}{\text{Kvs(Residual)}/(N-k)} \\ &= \frac{48.6/(6-4)}{75.6/(25-6)} = 6.11, \end{aligned}$$

vilket överstiger $F_{0.05}(k-l, N-k) = F_{0.05}(2, 19) = 3.5$. Vi kan alltså förkasta hollhypotesen att de genetiska faktorerna inte har någon signifikant inverkan på nivån 5%.

b) Vi räknar först ut den totala kvadratsumman

$$\begin{aligned} \text{Kvs(Total)} &= \text{Kvs(Avv från hypotes)} + \text{Kvs(Regr för hypotesmodell)} + \text{Kvs(Residual)} \\ &= 48.6 + 72.1 + 75.6 = 196.3. \end{aligned}$$

Hypotesmodellens förklaringsgrad blir då

$$R_1^2 = \frac{\text{Kvs(Regr för hypotesmodell)}}{\text{Kvs(Total)}} = \frac{72.1}{196.3} = 0.3673,$$

medan förklaringsgraden för grundmodellen är

$$\begin{aligned} R_0^2 &= \frac{\text{Kvs(Regr för grundmodell)}}{\text{Kvs(Total)}} = \frac{\text{Kvs(Avv från hypotes)+Kvs(Regr för hypotesmodell)}}{\text{Kvs(Total)}} \\ &= \frac{48.6+72.1}{196.3} = 0.6149. \end{aligned}$$

c) Vi räknar först ut skattningen av feltermernas varians utifrån grundmodellens residualer. Det följer av kalkylerna i a) att

$$\hat{\sigma}^2 = \text{Mkvs(Residual)} = \frac{75.6}{25-6} = 3.979.$$

Därefter bestämmer vi variansskattningen då man inte tar med några förklarande variabler i modellen. Den ges av

$$\hat{\sigma}_0^2 = \frac{\text{Kvs(Total)}}{25-1} = 8.1792.$$

Eftersom den efterfrågade justerade förklaringsgraden mäter variansreduktionen i grundmodellen, följer att

$$R^2_{0,\text{adj}} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = 1 - \frac{3.979}{8.179} = 0.5135.$$

Uppgift 3

a) Låt $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_{24})^T$ och $\hat{\hat{\boldsymbol{\mu}}} = (\hat{\hat{\mu}}_1, \ldots, \hat{\hat{\mu}}_{24})^T$ vara minsta kvadratskattningarna av observationernas väntevärdesvektor $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{24})^T$, under grund- respektive hypotesmodellen. Det följer av ledningen att

$$\text{Kvs(Avvikelse från hypotes)} = \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 = \text{Kvs(Monomer 3)} = 4.3,$$

eftersom monomer 3 är den variationskälla som ingår i grundmodellen men inte i hypotesmodellen. Eftersom grundmodellen har $k = 4$ parametrar, hypotesmodellen $l = 3$ parametrar och residualerna $N - k = 3 \cdot 8 - 4 = 20$ frihetsgrader, ger det en

$$\text{F-kvot} = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/(k-l)}{\text{Mkvs(Residual)}} = \frac{4.3}{28.2/20} = 3.05$$

som understiger $F_{0.05}(1, 20) = 4.35$. Vi kan alltså inte förkasta nollhypotesen att endast monomer 1 och monomer 2 påverkar plastens hållfasthet.

b) I första steget av BE utgår vi från grundmodellen i a), och testar den mot de tre olika hypotesmodeller som fås genom att ta bort en förklarande variabel (det vill säga monomer) i taget. Vi numrerar dessa tre hypotesmodeller enligt

Hypotesmodell 1: Monomer 2 och 3 som förklarande variabler,
Hypotesmodell 2: Monomer 1 och 3 som förklarande variabler,
Hypotesmodell 3: Monomer 1 och 2 som förklarande variabler.

I a) bestämdes F-kvoten då grundmodellen testas mot hypotesmodell 3. På motsvarande sätt fås

$$\text{F-kvot(Hypotesmodell 1)} = \frac{\text{Kvs(Monomer 1)}}{\text{Kvs(Residual)}/20} = \frac{7.1}{28.2/20} = 5.04$$

och

$$\text{F-kvot(Hypotesmodell 2)} = \frac{\text{Kvs(Monomer 2)}}{\text{Kvs(Residual)}/20} = \frac{6.2}{28.2/20} = 4.39.$$

Eftersom hypotesmodell 3 gav det minsta, och dessutom icke-signifikanta, värdet på F-kvoten, leder första steget av BE-schemat till att monomer 3 tas bort från grundmodellen.

c) I andra steget av BE-schemat har grundmodellen monomer 1 och 2 som förklarande variabler, det vill säga den modell som valdes i a). Vi testar denna grundmodell mot var och en av

Hypotesmodell 4: Monomer 1 som förklarande variabel,
Hypotesmodell 5: Monomer 2 som förklarande variabel.

Vi får nu att

$$\begin{aligned}\text{F-kvot(Hypotesmodell 4)} &= \frac{\text{Kvs(Monomer 2)}}{[\text{Kvs(Residual)}+\text{Kvs(Monomer 3)}]/21} \\ &= \frac{6.2}{(28.2+4.3)/21} = 4.01,\end{aligned}$$

eftersom monomer 2 ingår i grundmodellen i c) men inte i hypotesmodell 4, och residualdelen för grundmodellen i c) innehåller två variationskällor från den givna tabellen; residualdelen för grundmodellen i a) och b), samt monomer 3. På motsvarande sätt får vi att

$$\begin{aligned}\text{F-kvot(Hypotesmodell 5)} &= \frac{\text{Kvs(Monomer 1)}}{[\text{Kvs(Residual)}+\text{Kvs(Monomer 3)}]/21} \\ &= \frac{7.1}{(28.2+4.3)/21} = 4.58.\end{aligned}$$

Eftersom den minsta F-kvoten, för hypotesmodell 4, understiger $F_{0.05}(1, 21) = 4.32$, så stannar *inte* BE-schemat efter det andra steget. Istället väljs hypotesmodell 4, med monomer 1 som enda förklarande variabel, efter BE-schemats andra steg. BE-schemat fortsätter sedan med ett tredje steg, där det undersöks om den enda kvarvarande förklarande variabeln (monomer 1), efter BE-schemats andra steg, ska tas bort eller inte.

Uppgift 4

a) Vi kan skriva modellen som

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad 1 \leq i, j \leq 3, 1 \leq k \leq 2,$$

för hjärtfrekvensen hos patient k som tar medicin A på nivå i och medicin B på nivå j. Här anger $\mu = \sum_{i,j,k} E(Y_{ijk})/18$ det genomsnittliga väntevärdet för graden av hjärtklappning hos hela gruppen av patienter, α_i svarar mot effekten av medicin A, β_j mot effekten av medicin B och γ_{ij} mot samspelseffekten. Dessa parametrar uppfyller bivillkoren $\sum_i \alpha_i = \sum_j \beta_j = 0$,

samt $\sum_j \gamma_{ij} = \sum_i \gamma_{ij} = 0$. Dessa $7 = 1 + 1 + 5$ bivillkor ger totalt $9 = (1 + 3 + 3 + 9) - (1 + 1 + 5) = 9$ fria regressionsparametrar i modellen. Slutligen är $\varepsilon_{ijk} \sim N(0, \sigma^2)$ oberoende feltermer.

b) Antalet frihetsgrader för samspel och residual är $(3 - 1)(3 - 1) = 4$ respektive $3 \cdot 3 \cdot (2 - 1) = 9$. Det ger en

$$\text{F-kvot} = \frac{\text{Mkvs(Samspel)}}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Samspel)}/4}{\text{Kvs(Residual)}/9} = \frac{7.7/4}{9.3/9} = 1.86,$$

för att testa nollhypotesen att det inte finns något samspel ($H_0 : \gamma_{ij} \equiv 0$) mellan medicin A och B. Eftersom F-kvoten understiger $F_{0.05}(4, 9) = 3.63$, så kan vi inte förkasta nollhypotesen på signifikansnivån 5%.

c) Eftersom samspelet i b) inte var signifikant sätter vi samspelsparametrarna till noll. Det ger en ny kvadratsumma för variationskällan Residual som svarar mot summan av Kvs(Samspel) och Kvs(Residual), med totalt $4 + 9 = 13$ frihetsgrader. Eftersom försöket är balanserat är de två huvudeffekterna Medicin A och B ortogonala. Deras gemensamma effekt svarar därför mot en kvadratsumma som är summan av Kvs(Medicin A) och Kvs(Medicin B) i den givna tabellen. Denna variationskälla har alltså $(3 - 1) + (3 - 1) = 4$ frihetsgrader. För att testa nollhypotesen att varken Medicin A eller B har någon effekt ($H_0 : \alpha_i \equiv \beta_j \equiv 0$), så bildar vi således

$$\text{F-kvot} = \frac{\left[\text{Kvs(Medicin A)} + \text{Kvs(Medicin B)}\right]/4}{\left[\text{Kvs(Samspel)} + \text{Kvs(Residual)}\right]/13} = \frac{(9.1 + 8.4)/4}{(7.7 + 9.3)/13} = 3.35.$$

Eftersom $F_{0.05}(4, 13) = 3.18 < 3.35$, kan vi alltså förkasta nollhypotesen att de två medicinerna tillsammans inte har någon effekt på hjärtklappning, på nivån 5%.

Uppgift 5

a) Vi börjar med att bestämma γ_0. Från definitionen av en AR(1)-process följer att

$$\begin{aligned} \gamma_0 &= \text{Var}(X_t) \\ &= \text{Var}(\phi X_{t-1} + \varepsilon_t) \\ &= \phi^2 \text{Var}(X_{t-1}) + 2\phi \text{Cov}(X_{t-1}, \varepsilon_t) + \text{Var}(\varepsilon_t) \\ &= \phi^2 \gamma_0 + 2\phi \cdot 0 + \sigma_\varepsilon^2 \\ &= \phi^2 \gamma_0 + \sigma_\varepsilon^2, \end{aligned} \tag{5}$$

där vi i fjärde steget utnyttjade ledningen. Genom att lösa ut γ_0 ur (5) fås

$$\gamma_0 = \frac{\sigma_\varepsilon^2}{1 - \phi^2}. \tag{6}$$

För att bestämma γ_k för $k \geq 1$ används rekursion. Vi har att

$$\begin{aligned} \gamma_k &= \text{Cov}(X_t, X_{t+k}) \\ &= \text{Cov}(X_t, \phi X_{t+k-1} + \varepsilon_{t+k}) \\ &= \phi \text{Cov}(X_t, X_{t+k-1}) + \text{Cov}(X_t, \varepsilon_{t+k}) \\ &= \phi\gamma_{k-1} + 0 \\ &= \phi\gamma_{k-1}, \end{aligned} \tag{7}$$

där vi i näst sista steget utnyttjade ledningen. Genom att kombinera (6) med upprepad användning av (7) inses att

$$\gamma_k = \phi^k \phi_0 = \frac{\phi^k \sigma_\varepsilon^2}{1 - \phi^2} \tag{8}$$

för $k = 1, 2, \ldots$. För att bestämma autkorrelationsfunktionen så utnyttjas (8). Det ger

$$\rho_k = \text{Corr}(X_t, X_{t+k}) = \frac{\text{Cov}(X_t, X_{t+k})}{\sqrt{\text{Var}(X_t)}\sqrt{\text{Var}(X_{t+k})}} = \frac{\gamma_k}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_k}{\gamma_0} = \phi^k. \tag{9}$$

b) Eftersom en AR(1)-process är en Markovprocess gäller

$$\hat{X}_{T+k} = E(X_{T+k}|\boldsymbol{X}_T) = E(X_{T+k}|X_T). \tag{10}$$

Sedan kan vi antingen utnyttja $E(X_T) = E(X_{T+k}) = 0$, $\text{Var}(X_T) = \text{Var}(X_{T+k}) = \gamma_0$ och det faktum att (X_T, X_{T+k}) är tvådimensionellt normalfördelad, för att i kombination med (10) dra slutsatsen

$$\begin{aligned} \hat{X}_{T+k} &= E(X_{T+k}) + \frac{\sqrt{\text{Var}(X_{T+k})}}{\sqrt{\text{Var}(X_T)}} \text{Corr}(X_T, X_{T+k})(X_T - E(X_T)) \\ &= \text{Corr}(X_T, X_{T+k}) X_T = \phi^k X_T. \end{aligned}$$

Alternativt kan vi använda oss av definitionen av en AR(1)-process och skriva om X_{T+k} som

$$X_{T+k} = \phi^k X_T + \phi^{k-1}\varepsilon_{T+1} + \ldots \phi\varepsilon_{T+k-1} + \varepsilon_{T+k}. \tag{11}$$

Insättning av detta uttryck i (10) ger

$$\hat{X}_{T+k} = E\left(\phi^k X_T + \phi^{k-1}\varepsilon_{T+1} + \ldots \phi\varepsilon_{T+k-1} + \varepsilon_{T+k}|X_T\right) = \phi^k X_T, \tag{12}$$

eftersom alla feltermer $\varepsilon_{T+1}, \ldots, \varepsilon_{T+k}$ är oberoende av X_T.

c) Differensbildning av (11) och (12) ger ett prediktionsfel

$$X_{T+k} - \hat{X}_{T+k} = \phi^{k-1}\varepsilon_{T+1} + \ldots \phi\varepsilon_{T+k-1} + \varepsilon_{T+k} \tag{13}$$

av X_{T+k}. Genom att beräkna variansen av (13) får vi en geometrisk summa som förenklas till

$$\sigma_k^2 = \text{Var}(X_{T+k} - \hat{X}_{T+k}) = \phi^{2(k-1)}\sigma_\varepsilon^2 + \phi^2\sigma_\varepsilon^2 + \sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2(1 - \phi^{2k})}{1 - \phi^2}.$$

Notera att σ_k^2 är en strikt växande funktion av k med $\sigma_1^2 = \sigma_\varepsilon^2$ och $\lim_{k\to\infty} \sigma_k^2 = \gamma_0$.