STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

18 augusti 2021 8–15

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

Uppgift 1

a) Skriv den centrerade regressionsmodellen som

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 20,$$

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

$$\begin{array}{rcl} \hat{\tilde{\alpha}} & = & \sum_i Y_i/20 = 1580.0/20 = 79.0, \\ \hat{\beta} & = & \sum_i Y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 = 160.2/85.0 = 1.8847. \end{array} \tag{1}$$

Eftersom $\bar{x} = \sum_i x_i/20 = 60/20 = 3$ blir skattningen av den förväntade hjärtfrekvensen

$$\hat{\alpha} = \hat{\tilde{\alpha}} - \hat{\beta}\bar{x} = 79.0 - 1.8847 \cdot 3 = 73.346 = 73.3 \tag{2}$$

för personer som inte konsumerar kaffe.

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

$$\begin{array}{rcl} \mathrm{Var}(\hat{\alpha}) & = & \mathrm{Var}(\hat{\tilde{\alpha}}) + \mathrm{Var}(\hat{\beta}) \cdot \bar{x}^2 \\ & = & \frac{\sigma^2}{20} + \frac{\sigma^2 \bar{x}^2}{\sum_i (x_i - \bar{x})^2} \\ & = & \sigma^2 \left(\frac{1}{20} + \frac{3.0^2}{85.0} \right) \\ & = & 0.1559 \cdot \sigma^2. \end{array} \tag{3}$$

Antalet frihetsgrader för variationskällan Residual är 20-2=18. Av detta följer att

$$\hat{\sigma}^2 = \frac{\mathrm{Kvs(Residual)}}{18} = \frac{301.0}{18} = 16.72. \tag{4}$$

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

$$d = \sqrt{0.1559} \cdot \hat{\sigma} = \sqrt{0.1559 \cdot 16.72} = 1.6145.$$

c) Ett 95 % konfidensintervall för den förväntade hjärtslagsfrekvensen, för personer som inte dricker kaffe, är

$$\begin{aligned} I_\alpha &= (\hat{\alpha} - t_{0.025}(18) \cdot d, \hat{\alpha} + t_{0.025}(18) \cdot d) \\ &= (73.346 - 2.1009 \cdot 1.6145, 73.346 + 2.1009 \cdot 1.6145) \\ &= (69.95, 76.74), \end{aligned}$$

där värdet på t-kvantilen fås från tabell ($t_{0.025}(18) = \sqrt{F_{0.05}(1, 18)}$).

Uppgift 2

a) Vi börjar med att fylla i antalet frihetsgrader f i den fullständiga modellens variansanalystabell:

Variationskälla	f	Kvs
Vatten	1	4.5
Cement	1	4.4
Residual	22	22.0
Totalt	24	30.9

I första FS-steget testas två olika grundmodeller, med vatten respektive cement som förklarande variabel, med ett F-test mot en hypotesmodell som bara innehåller intercept. Eftersom variationskällan vatten har störst kvadratsumma räknar vi först ut F-kvoten för delmodellen som endast har vatten som förklarande variabel. Om vi använder denna delmodell som grundmodell får vi enligt ledningen

$$\begin{aligned} \text{Kvs(Regression)} &= \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 \\ &= \text{Kvs(Vatten)} \\ &= 4.5, \\ \text{Kvs(Residual)} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\ &= \text{Kvs(Cement)} + \text{Kvs(Residual)}_{\text{fullst}} \\ &= 4.4 + 22.0 \\ &= 26.4 \end{aligned}$$

där $\boldsymbol{Y}$ är observationsvektorn, medan $\hat{\boldsymbol{\mu}}$ och $\hat{\hat{\boldsymbol{\mu}}}$ är skattningar av dess väntevärde $\boldsymbol{\mu}$ enligt grund- respektive hypotesmodellen. Eftersom antalet frihetsgrader för Regression och Residual är 1 respektive $1 + 22 = 23$ får vi en

$$\begin{aligned} \text{F-kvot} &= \frac{\text{Kvs(Regression)}/1}{\text{Kvs(Residual)}/23} \\ &= \frac{4.5}{26.4/23} \\ &= 3.92, \end{aligned}$$

som understiger $F_{0.05}(1, 23) = 4.279$. Därför förkastas inte nollhypotesen, svarande mot att vattenmängd inte ger ett signifikant bidrag till betongens hållfasthet.

De andra delmodellen med bara cement som förklarande variabler har också 1 frihetsgrad för Regression och 23 frihetsgrader för Residual. Eftersom dess kvadratsumma Kvs(Regression) = Kvs(Cement) är lägre och deras kvadratsumma Kvs(Residual) = Kvs(Total) − Kvs(Regression) för residual är högre jämfört med modellen som bara har vatten som förklarande variabel, så kommer F-kvoten för denna delmodell vara lägre än 3.92. Därför förkastas inte nollhypotesen att ingen förklarande variabel ingår i detta test. Slutsatsen blir att FS-schemat stannar efter första steget och den valda modellen har endast intercept men inga förklarande variabler.

b) I första steget av BE-schemat testas den fullständiga modellen med två förklarande variabler mot två olika hypotesmodeller där antingen vatten eller cement tas bort som förklarande variabel. Vi börjar att testa grundmodellen mot den hypotesmodell M_1 som endast har vatten som förklarande variabel. Med hjälp av ledningen från a) får vi

$$\begin{aligned} \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 &= \text{Kvs(Regression)}_{\text{fullst}} - \text{Kvs(Regression)}_{M_1} \\ &= [\text{Kvs(vatten)} + \text{Kvs(cement)}] - \text{Kvs(vatten)} \\ &= \text{Kvs(cement)} \\ &= 4.4, \\ \text{Kvs(Residual)} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\ &= \text{Kvs(Residual)}_{\text{fullst}} \\ &= 22.0, \end{aligned} \tag{5}$$

och en

$$\begin{aligned} \text{F-kvot} &= \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/1}{\text{Kvs(Residual)}/22} \\ &= \frac{4.4}{22.0/22} \\ &= 4.4, \end{aligned} \tag{6}$$

som överstiger $F_{0.05}(1, 22) = 4.301$. Därför förkastas nollhypotesen att cement ska tas bort från grundmodellen. På motsvarande sätt fås en ännu större F-kvot (= 4.5/(22.0/22) = 4.5) då grundmodellen testas mot en hypotesmodell M_2 där endast cement finns med som förklarande variabel. Även här förkastas nollhypotesen, att vatten tas bort från grundmodellen. Slutsatsen blir att BE-schemat stannar efter första steget och den fullständiga modellen, med både vatten och cement som förklarande variabel, väljs.

BE-metoden väljer alltså en större modell än FS-metoden, och dess resultat är att föredra. Anledningen är att i FS-schemats första steg så inkluderas ingen av de två förklarande variablerna, eftersom dess effekter på responsvariabeln är för små i förhållande till residualernas storlek (som ju förutom den fullständiga modellens feltermer och skattningsfel även innehåller den förklarande variabel som inte testas med respektive F-kvot). Med andra ord finns det för mycket oförklarad variation i de två grundmodeller som testas i FS-schemats första steg, för att de effekter som testas ska bli signifikanta. Detta fenomen uppstår inte i BE-schemats första steg, eftersom residualerna

i F-testet härrör från den fullständiga modellen, och därmed inte innehåller någon oförklarad variation från vatten eller cement.

Uppgift 3

a) Modellen kan skrivas som

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{7}$$

för minskningen av ligninmängden hos träd $k \in \{1, 2\}$ inom den grupp av träd som injiceras med en koncentration på nivå $i \in \{1, 2, 3\}$ för svampsort 1 och på nivå $j \in \{1, 2, 3, 4, 5\}$ för svampsort 2. Parametern μ svarar mot medelvärdet av den förväntade ligninreduktionen för alla $3 \cdot 5 = 15$ grupper, α_i anger den systematiska effekten av då svampsort 1 har koncentration på nivå i, β_j den systematiska effekten då svampsort 2 har koncentration på nivå j. Slutligen anger γ_{ij} samspelet mellan de båda svampsorternas inverkan på ligninreduktionen. För att undvika överparametrisering inför vi totalt 9 linjärt oberoende bivillkor $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ (varav 1 bivillkor för α_i, 1 för β_j och $3 \cdot 5 - (3-1)(5-1) = 7$ för γ_{ij}). Feltermerna ε_i antas vara oberoende och normalfördelade med väntevärde 0 och varians σ^2.

b) För att testa grundmodellen (7) mot hypotesmodellen

$$H_0 : \gamma_{ij} = 0, \, \forall i, j$$

att det inte finns något samspel mellan de två svampsorternas inverkan på ligninreduktionen, bildar vi

$$\text{F-kvot} = \frac{\text{Mkvs(Samspel)}}{\text{Mkvs(Inom celler)}} = \frac{\text{Kvs(Samspel)}/8}{\text{Kvs(Inom celler)}/15} = \frac{15.2/8}{14.0/15} = 2.04.$$

Här utnyttjade vi att variationskällan Samspel har $(3-1)(5-1) = 8$ frihetsgrader, medan Inom celler har $3 \cdot 5(2-1) = 15$ frihetsgrader. Då F-kvoten har en $F(8, 15)$-fördelning under H_0 så jämför vi dess observerade värde med 95%-kvantilen

$$F_{0.05}(8, 15) = 2.64.$$

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

c) Då samspelet i b) inte var signifikant så väljer vi en additiv modell (=hypotesmodellen H_0 från b) ovan) som grundmodell. Alltså slår vi ihop de två variationskällorna Samspel och Inom celler till en ny variationskälla med 8+15=23 frihetsgrader. Vi skattar sedan feltermernas varians enligt

$$\hat{\sigma}^2 = \frac{\text{Kvs(Samspel)} + \text{Kvs(Inom celler)}}{8 + 15} = \frac{15.2 + 14.0}{23} = 1.2696.$$

Eftersom variationskällan Svampsort 1 har 3 nivåer och alltså 3-1=2 frihetsgrader får vi en

$$\text{F-kvot} = \frac{\text{Kvs(Svampsort 1)}/2}{\hat{\sigma}^2} = \frac{8.5/2}{1.2696} = 3.3475 < F_{0.05}(2, 23) = 3.422. \tag{8}$$

Således kan vi inte förkasta nollhypotesen H_0' att variation av den injicerade koncentrationen av svampsort 1 inte har någon effekt på trädets minskade ligninmängd, på nivån 5%. Om vi däremot *inte* hade inkluderat variationskällan Samspel i residualerna hade vi i stället fått en

$$\text{F-kvot} = \frac{\text{Kvs(Svampsort 1)}/2}{\text{Kvs(Inom celler)}/15} = \frac{8.5/2}{14.0/15} = 4.55 > F_{0.05}(2, 15) = 3.68, \tag{9}$$

som överstiger tröskelvärdet. Med andra ord hade nollhypotesen förkastats. Skillnaden mellan (8) och (9) illustrerar risken att inkludera ett icke-signifikant samspel i variationskällan residual.

Uppgift 4

a) Genom att utöka teckenschemat för det första fraktionella försöket med kolumner för enheten I och alla interaktioner av ordning 2 och 3 fås:

I	S	V	M	SV	SM	VM	SVM
+	-	-	-	+	+	+	-
+	+	-	+	-	+	-	-
+	-	+	-	-	+	-	+
+	+	+	+	+	+	+	+

Vi parar sedan ihop identiska kolumner och erhåller kopplingsmönstret $I = SM$, $S = M$, $V = SVM$, $SV = VM$. Alternativt ser man först att SM är kopplat till enheten och fyller sedan på med de andra tre kopplingarna enligt $S = SI = S(SM) = S^2M = IM = M$, $V = V(SM) = SVM$ och $SV = SV(SM) = VM$.

För det andra fraktionella försöket gör vi på motsvarande sätt. Utfyllnad av teckentabellen ger

I	S	V	M	SV	SM	VM	SVM
+	-	-	+	+	-	-	+
+	-	+	-	-	+	-	+
+	+	-	-	-	-	+	+
+	+	+	+	+	+	+	+

Genom att sammanföra identiska kolumner ser vi att kopplingsmönstret är $I = SVM$, $S = VM$, $V = SM$, $M = SV$. Alternativt kan vi först notera att SVM är kopplad till enheten I, och sedan bestämma de andra tre

kopplingarna utifrån det, t ex $S = SI = S(SVM) = S^2VM = IVM = VM$ osv.

b) I det första fraktionella försöket är huvudeffekterna S och M kopplade till varandra och kan därigenom inte särskiljas. För det andra fraktionella försöket tillhör de tre huvudeffekterna olika par av kopplade effekter. Varje huvudeffekt är alltså kopplad till en interaktionseffekt. Eftersom alla interaktionseffekter satts till 0 kan alla tre huvudeffekterna $\bar{S}$, $\bar{V}$ och $\bar{M}$ skattas för detta försök.

För att skatta huvudeffekterna för det andra fraktionella försöket inför vi observationsvektorn $\boldsymbol{Y} = (Y_{--+}, Y_{-+-}, Y_{+--}, Y_{+++})^T$, parametervektorn $\boldsymbol{\theta} = (\mu, \bar{S}, \bar{V}, \bar{M})^T$, och designmatrisen

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

som fås genom att till vänster om det givna teckenschemat addera en kolumn med ettor (svarande mot μ). Man kan sedan använda den allmänna formeln

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \frac{1}{4}\boldsymbol{A}^T\boldsymbol{Y}$$

för minsta kvadrat-skattningen av $\boldsymbol{\theta}$. Efter lite räkningar ser man att skattningarna av de tre huvudeffekterna blir

$$\begin{aligned} \hat{S} &= (-Y_{--+} - Y_{-+-} + Y_{+--} + Y_{+++})/4 = 2.25, \\ \hat{V} &= (-Y_{--+} + Y_{-+-} - Y_{+--} + Y_{+++})/4 = 1.25, \\ \hat{M} &= (Y_{--+} - Y_{-+-} - Y_{+--} + Y_{+++})/4 = 0.75. \end{aligned} \tag{10}$$

Alternativt kan man komma fram till (10) direkt genom att utgå från det andra försökets teckenschema, då dess kolumner är ortogonala. Eftersom antalet observationer är $N = 2^{3-1} = 4$ och antalet parametrar i $\boldsymbol{\theta}$ är $k = 4$, så är antalet frihetsgrader för residualrummet $N - k = 0$. Vi kan därför inte skatta feltermsvariansen σ^2.

c) Vi ställer upp en variansanalystabell för modellen i b) och använder ledningen för att räkna ut kvadratsummorna för alla variationskällor:

Variationskälla	f	Kvs
S	1	$4\hat{S}^2 = 20.25$
V	1	$4\hat{V}^2 = 6.25$
M	1	$4\hat{M}^2 = 2.25$
Totalt	3	28.75

För modellen i c) inkluderas faktor M i feltermerna. Det innebär att denna modell har samma variansanalystabell som i b), frånsett att variationskällan

M ändras till Residual. Eftersom variationskällan regression innehåller två faktorer S och V, följer av ledningen att

$$R^2 = \frac{\text{Kvs(Regression)}}{\text{Kvs(Total)}} = \frac{\text{Kvs}(S) + \text{Kvs}(V)}{\text{Kvs(Total)}} = \frac{20.25 + 6.25}{28.75} = 0.922,$$

eftersom de två variationskällorna S och V är ortogonala. Det innebär alltså att Kvs(Regression) för den additiva och tvåsidiga modellen i c) är summan av respektive faktors kvadratsumma.

Uppgift 5

a) Eftersom $Y_{ij} = \mu + \epsilon_{ij} = 1 \cdot \mu + \epsilon_{ij}$ så svarar μ mot ett intercept i modellen. Med andra ord är koefficienten 1 framför μ för alla observationer Y_{ij}. Därför ges designmatrisen av $\boldsymbol{A} = (1, \ldots, 1)^T = \mathbf{1}_N$, dvs en kolumnvektor med n stycken ettor. Vidare ges kovariansen mellan olika observationer av

$$\text{Cov}(Y_{ij}, Y_{kl}) = \text{Cov}(\epsilon_{ij}, \epsilon_{kl}) = \text{Cov}(\delta_i + \varepsilon_{ij}, \delta_k + \varepsilon_{kl}) = \begin{cases} \sigma_\delta^2 + \sigma_\epsilon^2, & i = k, j = l, \\ \sigma_\delta^2, & i = k, j \neq l, \\ 0, & i \neq k. \end{cases}$$

Av detta följer att feltermer ϵ_{ij} som härrör från olika nivåer i på faktorn är oberoende, vilket i sin tur medför att $\text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = \text{BDiag}(\boldsymbol{\Lambda}, \ldots, \boldsymbol{\Lambda})$ har en blockdiagonal struktur med matriser $\boldsymbol{\Lambda} = \sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^T + \sigma_\epsilon^2 \boldsymbol{I}_n$ längs blockdiagonalen. Här anger $\boldsymbol{I}_n$ identitetsmatrisenen av ordning n.

b) Låt $\boldsymbol{B} = (\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1}$. Eftersom $\hat{\mu} = \boldsymbol{B}\boldsymbol{Y}$ så följer att

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \boldsymbol{B}\text{Var}(\boldsymbol{Y})\boldsymbol{B}^T \\ &= (\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \boldsymbol{A} (\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \\ &= (\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} (\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} \\ &= (\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1}. \end{aligned}$$

c) Vi börjar med att notera att

$$\begin{aligned} \boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A} &= \mathbf{1}_N^T \boldsymbol{\Sigma}^{-1} \mathbf{1}_N \\ &= k \mathbf{1}_n^T \boldsymbol{\Lambda}^{-1} \mathbf{1}_n \\ &= k \mathbf{1}_n^T (\sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^T + \sigma_\epsilon^2 \boldsymbol{I}_n)^{-1} \mathbf{1}_n \\ &= k \mathbf{1}_n^T \left[-\sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^T / (n\sigma_\alpha^2 \sigma_\epsilon^2 + \sigma_\epsilon^4) + \boldsymbol{I}_n / \sigma_\epsilon^2 \right] \mathbf{1}_n \\ &= -kn^2 \sigma_\alpha^2 / (n\sigma_\alpha^2 \sigma_\epsilon^2 + \sigma_\epsilon^4) + kn/\sigma_\epsilon^2 \\ &= kn/(n\sigma_\alpha^2 + \sigma_\epsilon^2), \end{aligned} \tag{11}$$

där vi i fjärde ledet utnyttjade ledningen med $a = \sigma_\alpha^2$ och $b = \sigma_\epsilon^2$. Invertering av (11) ger att

$$\text{Var}(\hat{\mu}) = (\boldsymbol{A}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{A})^{-1} = \sigma_\alpha^2/k + \sigma_\epsilon^2/(kn). \tag{12}$$

På motsvarande sätt får vi att

$$\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1} = \mathbf{1}_N^T\boldsymbol{\Sigma}^{-1} = (\mathbf{1}_n^T\boldsymbol{\Lambda}^{-1}, \ldots, \mathbf{1}_n^T\boldsymbol{\Lambda}^{-1}), \tag{13}$$

där

$$\begin{aligned} \mathbf{1}_n^T\boldsymbol{\Lambda}^{-1} &= \mathbf{1}_n^T\left[-\sigma_\alpha^2\mathbf{1}_n\mathbf{1}_n^T/(n\sigma_\alpha^2\sigma_\epsilon^2+\sigma_\epsilon^4) + \boldsymbol{I}_n/\sigma_\epsilon^2\right] \\ &= \left[-n\sigma_\alpha^2/(n\sigma_\alpha^2\sigma_\epsilon^2+\sigma_\epsilon^4) + 1/\sigma_\epsilon^2\right]\mathbf{1}_n^T \\ &= \mathbf{1}_n^T/(n\sigma_\alpha^2+\sigma_\epsilon^2). \end{aligned} \tag{14}$$

Genom att sätta in (14) i (13) ser vi att

$$\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1} = \mathbf{1}_N^T/(n\sigma_\alpha^2+\sigma_\epsilon^2). \tag{15}$$

Därefter kombinerar vi (12) med (15) och drar slutsatsen att

$$\begin{aligned} \hat{\mu} &= (\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} \\ &= (\sigma_\alpha^2/k + \sigma_\epsilon^2/(kn))\mathbf{1}_N^T\boldsymbol{Y}/(n\sigma_\alpha^2+\sigma_\epsilon^2) \\ &= \mathbf{1}_N^T\boldsymbol{Y}/(kn) \\ &= \mathbf{1}_N^T\boldsymbol{Y}/N \\ &= (Y_{11}+\ldots+Y_{kn})/N \end{aligned}$$

är vanliga stickprovsmedelvärdet av alla observationer $\{Y_{ij}\}$.