STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

19 augusti 2020 9–17

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

Uppgift 1

a) Minsta kvadrat-skattningarna av α och β ges av

$$\begin{array}{rcl} \hat{\alpha} & = & \sum_i Y_i/20 = 210.0/20 = 10.5, \\ \hat{\beta} & = & \sum_i Y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2 = 202.5/45.0 = 4.50. \end{array} \tag{1}$$

b) Kalles förväntade blodtryckssänkning skattas till

$$\hat{\mu} = \hat{\alpha} + \hat{\beta} = 10.5 + 4.5 = 15.0. \tag{2}$$

De två skattningarna i (2) är oberoende stokastiska variabler. Detta medför att

$$\begin{array}{rcl} \mathrm{Var}(\hat{\mu}) & = & \mathrm{Var}(\hat{\alpha}_) + \mathrm{Var}(\hat{\beta}) \\ & = & \frac{\sigma^2}{20} + \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \\ & = & \sigma^2 \left(\frac{1}{20} + \frac{1}{45.0} \right) \\ & = & 0.0722 \cdot \sigma^2. \end{array} \tag{3}$$

Antalet frihetsgrader för att skatta feltermernas varians är 20-2=18. Av detta följer att

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{20} (Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2}{18} = \frac{105.0}{18} = 5.833. \tag{4}$$

Genom att först sätta in (4) i (3) och sedan ta kvadratroten ur det erhållna uttrycket, så erhålls medelfelet

$$d = \sqrt{0.0722} \cdot \hat{\sigma} = \sqrt{0.0722 \cdot 5.833} = 0.6491.$$

c) Ett 95 % konfidensintervall för Kalles förväntade blodtryckssänkning μ efter en vecka (enhet: mm Hg) är

$$\begin{array}{rcl} I_\mu & = & (\hat{\mu} - t_{0.025}(18) \cdot d, \hat{\mu} + t_{0.025}(18) \cdot d) \\ & = & (15.0 - 2.101 \cdot 0.6491, 15.0 + 2.101 \cdot 0.6491) \\ & = & (13.64, 16.36), \end{array}$$

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, med förkortningarna I, II och III för de tre ämnena:

Variationskälla	f	Kvs
I	1	8.0
II	1	6.0
III	1	3.0
Residual	26	23.0
Totalt	29	40.0

I BS-schemats första steg testas den fullständiga modellen $M_1 = (\mathrm{I}, \mathrm{II}, \mathrm{III})$, med alla de tre förklarande variablerna, som grundmodell, med ett F-test mot tre olika hypotesmodeller $M_2 = (\mathrm{I}, \mathrm{II})$, $M_3 = (\mathrm{I}, \mathrm{III})$ respektive $M_4 = (\mathrm{II}, \mathrm{III})$, där en förklarande variabel tagits bort i varje hypotesmodell. Vi börjar med att testa M_2 mot M_1. På grund av ortogonaliteten mellan de förklarande variablerna så får vi enligt ledningen att

$$\begin{aligned} \text{Kvs(Avvikelse från hypotes)} &= \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 \\ &= \text{Kvs(III)} \\ &= 3.0, \\ \text{Kvs(Residual)} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\ &= 23.0, \end{aligned} \tag{5}$$

där $\boldsymbol{Y}$ är observationsvektorn, och $\hat{\boldsymbol{\mu}}$, $\hat{\hat{\boldsymbol{\mu}}}$ skattningar av dess väntevärde enligt grund- respektive hypotesmodellen. Eftersom antalet frihetsgrader för Avvikelse från hypotes och Residual är 1 respektive 26 får vi en

$$\begin{aligned} \text{F-kvot} &= \frac{\text{Kvs(Avv. från hypotes)}/1}{\text{Kvs(Residual)}/26} \\ &= \frac{3.0}{23.0/26} \\ &= 3.39, \end{aligned} \tag{6}$$

som understiger $F_{0.05}(1, 26) = 4.225$. Därför förkastas inte nollhypotesen, svarande mot att variationer i III inte ger ett signifikant bidrag till plastens hållbarhet.

Om M_3 och M_4 testas mot M_1 fås kvadratsummor Kvs(Avv. från hyp) = Kvs(II) respektive Kvs(Avv. från hyp) = Kvs(I), som båda är större än värdet på Kvs(Avv. från hyp) = Kvs(III) i (5). I båda dessa tester har Avvikelse från hypotes 1 frihetsgrad och Residual 26 frihetsgrader, och Kvs(Residual) är dessutom densamma som i (6). Det gör att F-kvoterna då M_3 respektive M_4 testas mot M_1, båda är större än i (6).

Eftersom M_2 som hypotesmodell gav den minsta F-kvoten i första steget av BE-schemat, och M_2 inte förkastades i testet mot den fullständiga modellen M_1, så väljer vi M_2 efter första BS-steget.

b) I andra steget av BS testas $M_2 = (\mathrm{I}, \mathrm{II})$ som grundmodell mot två olika hypotesmodeller, $M_5 = (\mathrm{I})$ respektive $M_6 = (\mathrm{II})$. Vi börjar att testa M_5 som hypotesmodell mot M_2. Enligt ledningen i a) och b) så får vi, på grund av ortogonaliteten mellan prediktorerna, att

$$\begin{aligned} \text{Kvs(Avv. från hypotes)} &= \|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2 \\ &= \text{Kvs(II)} \\ &= 6.0, \\ \text{Kvs(Residual)}_{M_2} &= \|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2 \\ &= \text{Kvs(III)} + \text{Kvs(Residual)}_{M_1} \\ &= 3.0 + 23.0 \\ &= 26.0, \end{aligned} \tag{7}$$

och en

$$\begin{aligned} \text{F-kvot} &= \frac{\text{Kvs(Avv. från hypotes)}/1}{\text{Kvs(Residual)}_{M_2}/(1+26)} \\ &= \frac{6.0}{26.0/27} \\ &= 6.23, \end{aligned} \tag{8}$$

som överstiger $F_{0.05}(1, 27) = 4.21$. Därför förkastas nollhypotesen M_5, svarande mot att II inte tas bort från modellen. Motsvarande test mellan hypotesmodell M_6 och grundmodell M_2 ger en F-kvot vars täljare Kvs(Avv. från hyp) = Kvs(I) = 8.0 är större än i (8), medan F-kvotens nämnare är densamma som i (8). Det gör att F-kvoten då M_6 testas mot M_2 är större än i (8).

Eftersom testet mellan M_5 och M_2 gav den minsta F-kvoten i BS-schemats andra steg, och hypotesmodellen M_5 förkastades, så stannar BS-schemat i andra steget. Den slutgiltigt valda modellen är alltså $M_2 = (\mathrm{I}, \mathrm{II})$, där Ämne 1 och Ämne 2, men inte Ämne 3, ingår som förklarande variabler.

Uppgift 3

a) Modellen är en tvåsidig variansanalys typ I med sampspel. Därför bör varje cell i, j ha en egen parameter $E(Y_{ijk}) = \mu_{ij}$. Det ger totalt $4 \times 3 = 12$ parametrar. I den givna modellen $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ har vi 1 parameter μ för den genomsnittliga koncentrationen av antikroppar hos alla personer i undersökningen, 4 åldersparametrar α_i, 3 regionsparametrar β_j samt $4 \times 3 = 12$ samspelsparametrar γ_{ij}. Således har vi totalt $1 + 4 + 3 + 12 = 20$ regressionsparametrar, vilket är $20 - 12 = 8$ för många. Vi bör därför införa 8 oberoende linjära restriktioner på de 20 parametrarna, nämligen

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 &= 0, \\ \beta_1 + \beta_2 + \beta_3 &= 0, \\ \textstyle\sum_{j=1}^{3} \gamma_{ij} &= 0, \quad i = 1, \ldots, 4, \\ \textstyle\sum_{i=1}^{4} \gamma_{ij} &= 0, \quad j = 1, 2. \end{aligned}$$

Notera att även $\sum_{i=1}^{4} \gamma_{i3} = 0$, men denna linjära restriktion är linjärt beroende av de som angivits ovan. Av de kvarvarande 12 fritt varierande

parametrarna svarar 1 parameter μ mot den genomsnittliga koncentrationen av antikroppar och de övriga $12-1 = 11 = (4-1)+(3-1)+(4-1)(3-1)$ mot det totala antalet frihetsgrader för variationskällorna Ålder, Region och Samspel.

b) Antalet frihetsgrader för variationskällan Samspel är $(4-1)(3-1) = 6$, och för Inom celler är antalet frihetsgrader $4 \cdot 3(3-1) = 24$. För att testa nollhypotesen att det inte finns ett samspel mellan ålder och region vad bildandet av antikroppar, så bildar vi en

$$\text{F-kvot} = \frac{\text{Mkvs(Samspel)}}{\text{Mkvs(Inom celler)}} = \frac{\text{Kvs(Samspel)}/6}{\text{Kvs(Inom celler)}/24} = \frac{9.5/6}{25.0/24} = 1.52.$$

Då denna F-kvot inte överstiger tröskelvärdet $F_{0.05}(6, 24) = 2.51$, så kan vi inte förkasta nollhypotesen att samspel saknas på signifikansnivån 5%.

c) Eftersom samspelet i b) inte var signifikant så inkluderar vi denna variationskälla i skattningen av feltermernas varians σ^2, svarande mot att sätta alla samspelsparametrar γ_{ij} till 0. Det ger en skattning

$$\hat{\sigma}^2 = \frac{\text{Kvs(Samspel)} + \text{Kvs(Residual)}}{6 + 24} = \frac{9.5 + 25}{30} = 1.58$$

av σ^2 med totalt $6 + 24 = 30$ frihetsgrader. Motsvarande tvåsidiga 95% konfidensinterfall för standardavvikelsen σ blir

$$\begin{aligned} I_\sigma &= \left(\hat{\sigma}\sqrt{\tfrac{30}{\chi^2_{0.025}(30)}}, \hat{\sigma}\sqrt{\tfrac{30}{\chi^2_{0.975}(30)}}\right) \\ &= \left(\sqrt{\tfrac{1.58 \cdot 30}{46.98}}, \sqrt{\tfrac{1.58 \cdot 30}{16.79}}\right) \\ &= (1.00, 1.68), \end{aligned}$$

där vi i andra steget utnyttjade ledningen.

Uppgift 4

a) Vi börjar med att komplettera den givna tabellen genom att räkna ut de fyra cellmedelvärdena $\bar{Y}_{ij\cdot} = (Y_{ij1} + Y_{ij2})/2$, samt totalmedelvärdet $\bar{Y}_{\cdots}$:

T	S	Y_{ij1}	Y_{ij2}	$\bar{Y}_{ij\cdot}$
-	-	3.7	4.1	3.9
+	-	4.9	4.5	4.7
-	+	5.3	5.5	5.4
+	+	6.5	6.3	6.4
Medel				5.1

Minsta kvadrat-skattningar av de två huvudeffekterna, och av samspelseffekten, ges av

$$\begin{aligned} \hat{T} &= (-\bar{Y}_{--\cdot} + \bar{Y}_{+-\cdot} - \bar{Y}_{-+\cdot} + \bar{Y}_{++\cdot})/4 = 0.45, \\ \hat{S} &= (-\bar{Y}_{--\cdot} - \bar{Y}_{+-\cdot} + \bar{Y}_{-+\cdot} + \bar{Y}_{++\cdot})/4 = 0.80, \\ \widehat{TS} &= (+\bar{Y}_{--\cdot} - \bar{Y}_{+-\cdot} - \bar{Y}_{-+\cdot} + \bar{Y}_{++\cdot})/4 = 0.05. \end{aligned} \tag{9}$$

Alternativt kan man formulera den reducerade modellen med cellmedelvärden som en multipel linjär regressionsmodell där $\boldsymbol{Y} = (\bar{Y}_{--\cdot}, \bar{Y}_{+-\cdot}, \bar{Y}_{-+\cdot}, \bar{Y}_{++\cdot})^T$ är observationsvektor, $\boldsymbol{\theta} = (\mu, \bar{T}, \bar{S}, \overline{TS})^T$ parametervektor och

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

designmatris. Det ger en minsta kvadrat-skattning $\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \boldsymbol{A}^T\boldsymbol{Y}/4$ av parmetervektorn. Notera sedan att de tre sista komponenterna av $\hat{\boldsymbol{\theta}}$ överensstämmer med de tre skattningarna i (9).

b) Vi har att

$$\begin{aligned} \text{Kvs(Inom celler)} &= \textstyle\sum_{i,j}[(Y_{ij1} - \bar{Y}_{ij\cdot})^2 + (Y_{ij1} - \bar{Y}_{ij\cdot})^2] \\ &= \tfrac{1}{2}\textstyle\sum_{i,j}(Y_{ij1} - Y_{ij2})^2 \\ &= \tfrac{1}{2}[(3.7-4.1)^2 + (4.9-4.5)^2 + (5.3-5.5)^2 + (6.5-6.3)^2] \\ &= 0.2. \end{aligned}$$

Vi noterar att antalet frihetsgrader för variationskällan Inom celler är $2 \cdot 2(2-1) = 4$. Därav följer att

$$\hat{\sigma}^2 = \text{Mkvs(Inom celler)} = \frac{\text{Kvs(Inom celler)}}{4} = 0.05$$

är en väntevärdesriktig skattning av feltermsvariansen σ^2.

c) Observera att alla cellmedervärden har varians $\text{Var}(\bar{Y}_{ijk}) = \sigma^2/2$. Parameterskattningen $\hat{\boldsymbol{\theta}}$ i a) har därför kovariansmatrisen $\text{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2/2 \cdot (\boldsymbol{A}^T\boldsymbol{A})^{-1} = \sigma^2\boldsymbol{I}_4/8$, där $\boldsymbol{I}_4$ är identitetsmatrisen av ordning 4. Eftersom $\text{Var}(\hat{T})$ är det andra diagonalelementet i denna kovariansmatris följer att $\text{Var}(\hat{T}) = \sigma^2/8$. Detta kan också erhållas direkt från (9), genom

$$\text{Var}(\hat{T}) = \frac{1}{4^2}\left(\text{Var}(\bar{Y}_{--\cdot}) + \text{Var}(\bar{Y}_{+-\cdot}) + \text{Var}(\bar{Y}_{-+\cdot}) + \text{Var}(\bar{Y}_{++\cdot})\right) = \frac{4 \cdot \sigma^2/2}{16} = \frac{\sigma^2}{8}.$$

Medelfelet för skattningen är alltså

$$d = \sqrt{\frac{\hat{\sigma}^2}{8}} = \sqrt{\frac{0.05}{8}} = 0.0791.$$

Eftersom variationskällan Inom celler enligt b) har 4 frihetsgrader så följer att ett 95% konfidensintervall för $\bar{T}$ ges av

$$\begin{aligned} I &= (\hat{T} - t_{0.025}(4)d, \hat{T} + t_{0.025}(4)d) \\ &= (0.45 - 2.776 \cdot 0.0791, 0.45 + 2.776 \cdot 0.0791) \\ &= (0.230, 0.670), \end{aligned}$$

där $t_{0.025}(4) = \sqrt{F_{0.05}(1,4)}$ fås ur tabell. Eftersom 0 ligger utanför intervallet har stektiden en signifikant inverkan på kundnöjdheten, på nivån 5%.

Uppgift 5

a) Låt $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ vara observtionsvektorn och

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{21} \\ \vdots & \vdots \\ x_{1N} & x_{2N} \end{pmatrix}$$

den del av designmatrisen $\boldsymbol{A}$ som härrör från de två förklarande variablerna. Skattningen av de två effektparametrarna $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ ges av $\hat{\boldsymbol{\beta}} = \boldsymbol{S}^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, där

$$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}$$

och $s_{jk} = \sum_{i=1}^{N}(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)$. Variansmatrisen för skattningen av effektparametrarna är $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{S}^{-1}$. Från det översta diagonalelementet av denna variansmatris erhålls $\mathrm{Var}(\hat{\boldsymbol{\beta}}_1) = \sigma^2(\boldsymbol{S}^{-1})_{11}$ Om β_2 vore känd skulle variansen för skattningen av β_1 ges av $\mathrm{Var}_0(\hat{\boldsymbol{\beta}}_1) = \sigma^2 s_{11}^{-1}$. Variationsinflationsfaktorn (VIF) anger den relativa ökningen av variansen för skattningen av β_1, på grund av att vi även måste skatta β_2. Det svarar mot

$$\mathrm{VIF} = \frac{\mathrm{Var}(\hat{\beta}_1)}{\mathrm{Var}_0(\hat{\beta}_1)} = \frac{\sigma^2(\boldsymbol{S}^{-1})_{11}}{\sigma^2 s_{11}^{-1}} = s_{11}(\boldsymbol{S}^{-1})_{11}. \tag{10}$$

b) Vi ansätter en enkel linjär regressionsmodell

$$x_{1i} = \eta + \gamma(x_{2i} - \bar{x}_2) + \epsilon_i, \quad i = 1, \ldots, N, \tag{11}$$

för att förklara x_{1i} med hjälp av x_{2i}. Låt $\hat{\eta} = \bar{x}_1$ och $\hat{\gamma} = s_{12}/s_{22}$ vara minsta kvadrat-skattningarna av η och γ, samt $\hat{x}_{1i} = \hat{\eta} + \hat{\gamma}(x_{2i} - \bar{x}_2)$ den prediktion av x_{1i} som erhålls från regressionsmodellen (11). Den sökta förklaringsgraden är

$$R_1^2 = \frac{\sum_{i=1}^{N}(\hat{x}_{1i} - \bar{x}_1)^2}{\sum_{i=1}^{N}(x_{1i} - \bar{x}_1)^2} = \frac{\hat{\gamma}^2 s_{22}}{s_{11}} = \frac{s_{12}^2}{s_{11}s_{22}}. \tag{12}$$

c) Med hjälp av ledningen så kan variationsinflationsfaktorn i (10) skrivas om enligt

$$\mathrm{VIF} = s_{11} \cdot \frac{s_{22}}{s_{11}s_{22} - s_{12}^2} = \frac{s_{11}s_{22}}{s_{11}s_{22} - s_{12}^2}. \tag{13}$$

Vidare följer av (12) och (13) att

$$\frac{1}{1 - R_1^2} = \frac{s_{11}s_{22}}{s_{11}s_{22} - s_{12}^2} = \mathrm{VIF}.$$