STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

25 november 2020 9–16

Examinator: Ola Hössjer. Kan nås under skrivtiden via mobil (070/672 12 18) eller mejl (ola@math.su.se).

Inlämning: Lösningar mejlas till examinator senast kl 16 i form av en pdf-fil. Denna fil kan antingen innehålla inscannade och handskrivna lösningar eller lösningar som skrivits ned i en ordbehandlare (t ex LaTeX).

Återlämning: Meddelas via kurshemsidan, webbaserat kursforum eller per mejl.

Tillåtna hjälpmedel: Miniräknare och formelsamling, samt lärobok och andra skriftliga informationskällor. Tabell över F-kvantiler återfinns nedan. Det gäller även att $\chi^2_{0.05}(1) \approx 3.8$. Det är inte tillåtet att ta hjälp av andra personer.

Resonemang skall vara tydliga och lätta att följa. Varje korrekt och fullständigt löst uppgift ger 10 poäng. Följande gränser gäller för betygen A-E:

A	B	C	D	E
45	40	35	30	25

Uppgift 0

Skriv en försäkran att du löst alla uppgifter självständigt. Detta krävs för att tentan ska rättas.

(0 p)

Uppgift 1

Vid ett läkemedelsföretag har man utvecklat ett nytt vaccin som visat sig ha biverkningar i form av en inflammation, som uppstår alldeles efter att vaccinet injicerats. En grupp biostatistiker vid läkemedelsföretaget har

fått i uppgift att uppskatta hur stor denna effekt är genom att bestämma snabbsänkan CRP=Y (enhet: mg/L) hos individer som fått dosen x mg av vaccinet. Man undersöker 20 personer i en blind studie där deltagarna slumpmässigt delas in i fem lika stora grupper, där dosen 0, 1, 2, 3, och 4 mg ges till personerna i respektive grupp. Därefter ställer man upp en enkel linjär regressionsmodell

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

för CRP hos individ i med dosen $x_i \in \{0, 1, 2, 3, 4\}$, och där $\bar{x} = \sum_{i=1}^{20} x_i/20$ anger genomsnittlig dos. Vidare antas ε_i vara oberoende och $N(0, \sigma^2)$-fördelade feltermer. Resultatet av undersökningen sammanfattas i följande tabell:

x	Antal ind.	Medel	Stvar
0	4	2.0	1.0
1	4	3.4	1.8
2	4	4.0	2.4
3	4	5.2	2.0
4	4	6.1	2.8
Total	20	20.7	10.0

Här anger Medel och Stvar stickprovsmedelvärdet respektive stickprovsvariansen av snabbsänkan i respektive grupp. Syftet med undersökningen är att uppskatta värdet på β, dvs hur starka biverkningar vaccinet har.

a) Beräkna minsta kvadrat-skattningen $\hat{\beta}$ av β. (Ledning: Använd tabellen för att först beräkna relevanta summor med avseende på individer $i = 1, \ldots, 20$.) (3 p)

b) Beräkna $\mathrm{Var}(\hat{\beta})$, uttryckt i σ^2. (2 p)

c) För att skatta σ^2 vill man *inte* använda sig av residualkvadratsumman från regressionsanalysen. Man förmodar att regressionsmodellen (1) är något för enkel, så att residualerna fångar upp ett icke-linjärt samband mellan $E(Y_i)$ och x_i. Istället använder man stickprovsvarianserna ovan för att skatta σ^2. Beräkna denna skattning av σ^2 och använd sedan b) för att beräkna medelfelet $\widehat{\mathrm{Var}}(\hat{\beta})^{1/2}$ för skattningen av β. (3 p)

d) Beräkna ett tvåsidigt konfidensintervall för β med konfidensgrad 95%, där du använder medelfelet från c). Har vaccinet någon signifikant biverkning? (2 p)

Uppgift 2

Kalle vill mäta längderna θ_1, θ_2 och θ_3 på tre pappersark (enhet: cm). Han genomför ett försök med sex mätningar. I de tre första mätningarna lägger han två av pappersarken efter varandra på bordsytan och mäter totallängden

(exempelvis $\theta_1 + \theta_2$). I de tre sista mätningarna lägger han två av pappersarken efter varandra på bordsytan och det tredje arket ovanpå, och mäter skillnaden mellan den långa sträckan med två ark och den korta sträckan med ett ark (exempelvis mäts $\theta_1 + \theta_2 - \theta_3$ om ark 3 läggs överst). Mätfelen i alla mätningar antas oberoende och normalfördelade med väntevärde 0 och varians σ^2. Resultatet av Kalles försök sammanfattas i nedanstående tabell, där koefficienten för varje ark är 1 om det ligger på bordsytan och -1 om det ligger på ovanpå ett annat ark.

Mätning	Ark 1	Ark 2	Ark 3	Mätresultat
1	1	1	0	57.4
2	1	0	1	58.4
3	0	1	1	58.4
4	-1	1	1	30.2
5	1	-1	1	28.7
6	1	1	-1	27.7

a) Ställ upp en allmän linjär modell för försöket och beräkna sedan minsta kvadrat-skattningarna $\hat{\theta}_1$, $\hat{\theta}_2$ och $\hat{\theta}_3$ av de tre arkens längder. (4 p)
b) Låt $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ beteckna parametervektorn. Kalle vill testa nollhypotesen $H_0 : \theta_1 = \theta_2 = \theta_3$ att alla arken är lika långa. Formulera denna hypotes på formen $\boldsymbol{\theta} = \boldsymbol{B}\lambda$ för lämpligt vald matris $\boldsymbol{B}$ och tal λ. (2 p)
c) Man kan visa att Kvs(Residual) = 0.30. Använd denna information och data från tabellen ovan för att testa nollhypotesen från b) på nivån 5% med ett F-test. (Ledning: Börja med att först beräkna en minsta kvadratskattning $\hat{\lambda}$ av λ under H_0 med hjälp av designmatrisen $\boldsymbol{C}$ för hypotesmodellen (som är en funktion av $\boldsymbol{B}$ samt designmatrisen $\boldsymbol{A}$ för grundmodellen). Skattningen av $\boldsymbol{\mu} = \boldsymbol{A\theta}$ under grund- och hypotesmodellen ges av $\hat{\boldsymbol{\mu}} = \boldsymbol{A}\hat{\boldsymbol{\theta}}$ respektive $\hat{\hat{\boldsymbol{\mu}}} = \boldsymbol{C}\hat{\lambda}$.) (4 p)

Uppgift 3

I en fabrik framställs en viss typ av blå tapetvåder. Man har upptäckt att färgtonen skiljer sig åt mellan olika serier, dels på grund av varierande tjocklek på våden och dels för att färgnyansen varierar i färgbadet som våden doppas i. För att avgöra vilken inverkan dessa två faktorer har anställs en statistiker. Hon genomför ett försök enligt en tvåvägs variansanalys typ II, med tjocklek och färg som slumpmässiga faktorer. Totalt framställs fyra olika tapetvåder, med olika tjocklekar, som var och en doppas i tre olika färgbad. För varje serie (kombination av tjocklek och färgbad) genomförs två mätningar. Försöket sammanfattas med modellen

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

för färgnyansen hos mätning nummer k för serien med tjocklek i och färgnyans j. Här anger $\mu = E(Y_{ijk})$ den genomsnittliga färgnyansen, medan $\{\alpha_i\}_{i=1}^4$,

$\{\beta_j\}_{j=1}^3$, $\{\gamma_{ij}\}_{i,j}$ och $\{\varepsilon_{ijk}\}_{ijk}$ är oberoende och normalfördelade stokastiska variabler med väntevärde 0 och varianser σ_α^2, σ_β^2, σ_γ^2 respektive σ_ϵ^2.

a) Resultatet av variansanalysen sammanfattas i fölande variansanalystabell:

Variationskälla	Kvs	f	Mkvs
Tjocklek	18.0		
Färg	18.4		
Samspel	7.2		
Inom celler	6.0		
Totalt	52.8		

Fyll i de saknade uppgifterna i tabellen. Använd sedan (utan bevis) formler för $E[\text{Mkvs}(V)]$ för olika variationskällor V för att skatta de fyra variankomponenterna σ_ϵ^2, σ_γ^2, σ_α^2 och σ_β^2. (5 p)

b) Låt $\hat{\mu} = \bar{Y}_{\cdots}$ vara en skattning av $\hat{\mu}$. Ge ett uttryck för $\text{Var}(\hat{\mu})$ som är en linjärkombination av de olika varianskomponenterna. Använd sedan resultatet från a) för att beräkna medelfelet $\widehat{\text{Var}}(\hat{\mu})^{1/2}$ för skattningen av μ. (5 p)

Uppgift 4

Vid en ortopedklinik genomförs knäoperationer för patienter med artros. Man vill undersöka hur rehabiliteringstiden (dvs den tid i månader det tar tills patienten kan gå själv efter operationen) varierar med storleken på knäprotesen (faktor S) samt mängden skruvar man använder för att fästa protesen (faktor M). För att få svar på denna fråga utförs ett 2^2-försök med två replikat, där båda faktorerna varieras på en låg nivå - (svarande mot -1) och en hög nivå + (svarande mot 1). Rehabiliteringstiden för patient $k \in \{1, 2\}$ av de som har storlek på nivån $i \in \{-, +\}$ och mängden skruvar på nivån $j \in \{-, +\}$ antas följa modellen

$$Y_{ijk} = \mu + \bar{S} \cdot i + \bar{M} \cdot j + \overline{SM} \cdot ij + \varepsilon_{ijk},$$

där feltermerna $\varepsilon_{ijk} \sim N(0, \sigma^2)$ är oberoende. Resultatet av undersökningen framgår av följande tabell:

S	M	Y_{ij1}	Y_{ij2}
-	-	6.4	6.0
+	-	8.0	7.6
-	+	5.6	5.8
+	+	7.0	7.2

a) Beräkna minsta kvadrat-skattningar $\hat{S}$ och $\hat{M}$ av de två huvudeffekterna, samt $\widehat{SM}$ av samspelseffekten, genom att först bilda cellmedelvärden $\bar{Y}_{ij\cdot}$. (3 p)

b) Beräkna kvadratsumman inom celler och därefter en väntevärdesriktig skattning av σ^2. (3 p)

c) Bestäm ett 95% konfidensintervall för den förväntade skillnaden $2\bar{M}$ i rehabiliteringstid, hos en patient med slumpmässigt vald protes i, mellan två scenarier där ett större respektive mindre antal skruvar ($j = +$ och $j = -$) används för att fästa protesen. Är huvudeffekten för denna faktor signifikant? (4 p)

Uppgift 5

Låt $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ vara ett dataset som beskrivs av den allmänna linjära modellen

$$\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon}, \tag{2}$$

där $\boldsymbol{A}$ är en $N \times k$ designmatris och $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T$ en vektor med oberoende och $N(0, \sigma^2)$-fördelade feltermer.

a) Hattmatrisen $\boldsymbol{H} = (h_{ij})_{i,j=1}^N$ projicerar observationsvektorn $\boldsymbol{Y}$ ned på det delrum av R^N som spänns upp av grundmodellen (2). Definiera $\boldsymbol{H}$ med hjälp av designmatrisen $\boldsymbol{A}$. (1 p)

b) Låt $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}} = \boldsymbol{Y} - \boldsymbol{A}\hat{\boldsymbol{\theta}}$ vara residualvektorn, där $\hat{\boldsymbol{\theta}}$ är minsta kvadratskattningen av parametervektorn. Låt vidare

$$\text{Kvs(Residual)} = \|\boldsymbol{e}\|^2 = \sum_{i=1}^N e_i^2$$

vara residualkvadratsumman. Uttryck residualvektorn $\boldsymbol{e}$ som funktion av hattmatrisen $\boldsymbol{H}$ och feltermsvektorn $\boldsymbol{\varepsilon}$. Använd detta för att bestämma kovariansmatrisen $\text{Var}(\boldsymbol{e})$ och sedan visa att

$$E[\text{Kvs(Residual)}] = (N - k)\sigma^2.$$

(Ledning: Du får utan bevis använda att $\sum_{i=1}^N h_{ii} = k$.) (4 p)

c) Låt

$$\boldsymbol{Z} = \boldsymbol{A\theta} + \boldsymbol{\epsilon}$$

vara ett nytt dataset med samma designmatris och parametervektor som för $\boldsymbol{Y}$, men med en ny feltermsvektor $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ som är oberoende av feltermerna ε_i i det ursprungliga datamaterialet. Vidare antas ϵ_i vara sinsemellan oberoende och $N(0, \sigma^2)$-fördelade. Använd vektorn $\hat{\boldsymbol{\mu}} = \boldsymbol{A}\hat{\boldsymbol{\theta}}$ för att prediktera $\boldsymbol{Z}$ och låt

$$\text{Kvs(Prediktion)} = \|\boldsymbol{Z} - \hat{\boldsymbol{\mu}}\|^2 = \sum_{i=1}^N (Z_i - \hat{\mu}_i)^2$$

vara kvadratsumman för prediktionsfelen. Uttryck prediktionsfelsvektorn $\boldsymbol{Z} - \hat{\boldsymbol{\mu}}$ med hjälp av $\boldsymbol{H}$, $\boldsymbol{\varepsilon}$ och $\boldsymbol{\epsilon}$. Använd sedan detta för att visa att

$$E[\text{Kvs(Prediktion)}] = (N + k)\sigma^2. \tag{3}$$

(3 p)

d) Anta att vi inte har tillgång till $\boldsymbol{Z}$ och vill uppskatta det förväntade kvadratiska prediktionsfelet (3) med hjälp av det första datasetets residualkvadratsumma Kvs(Residual). Eftersom residualkvadratsumman kommer att underskatta prediktionsfelens kvadratsumma så ansätter vi

$$\widehat{\text{Kvs}}(\text{Prediktion}) = \text{Kvs(Residual)} + C\hat{\sigma}^2, \tag{4}$$

där

$$\hat{\sigma}^2 = \text{Mkvs(Residual)} = \frac{\text{Kvs(Residual)}}{N - k}$$

är en skattning av feltermvariansen σ^2 med hjälp av det första datamaterialets kvadratsumma. Hur bör C väljas i (4) för att $\widehat{\text{Kvs}}(\text{Prediktion})$ ska vara en väntevärdesriktig skattning av Kvs(Prediktion)? Ange kort vilken relevans (4), med det valda värdet på C, kan ha som modellvalskriterium. (2 p)

	$f_1 = 1$	2	3	4	5	6	7	8	9	10
$f_2 = 1$	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
3	10.1	9.6	9.3	9.1	9.0	8.9	8.9	8.8	8.8	8.8
4	7.7	6.9	6.6	6.4	6.3	6.2	6.1	6.0	6.0	6.0
5	6.6	5.8	5.4	5.2	5.1	5.0	4.9	4.8	4.8	4.7
6	6.0	5.1	4.8	4.5	4.4	4.3	4.2	4.1	4.1	4.1
7	5.6	4.7	4.3	4.1	4.0	3.9	3.8	3.7	3.7	3.6
8	5.3	4.5	4.1	3.8	3.7	3.6	3.5	3.4	3.4	3.3
9	5.1	4.3	3.9	3.6	3.5	3.4	3.3	3.2	3.2	3.1
10	5.0	4.1	3.7	3.5	3.3	3.2	3.1	3.1	3.0	3.0
11	4.8	4.0	3.6	3.4	3.2	3.1	3.0	2.9	2.9	2.9
12	4.7	3.9	3.5	3.3	3.1	3.0	2.9	2.8	2.8	2.8
13	4.7	3.8	3.4	3.2	3.0	2.9	2.8	2.8	2.7	2.7
14	4.6	3.7	3.3	3.1	3.0	2.8	2.8	2.7	2.6	2.6
15	4.5	3.7	3.3	3.1	2.9	2.8	2.7	2.6	2.6	2.5
16	4.5	3.6	3.2	3.0	2.9	2.7	2.7	2.6	2.5	2.5
17	4.5	3.6	3.2	3.0	2.8	2.7	2.6	2.5	2.5	2.4
18	4.4	3.6	3.2	2.9	2.8	2.7	2.6	2.5	2.5	2.4
19	4.4	3.5	3.1	2.9	2.7	2.6	2.5	2.5	2.4	2.4
20	4.4	3.5	3.1	2.9	2.7	2.6	2.5	2.4	2.4	2.3
21	4.3	3.5	3.1	2.8	2.7	2.6	2.5	2.4	2.4	2.3
22	4.3	3.4	3.0	2.8	2.7	2.5	2.5	2.4	2.3	2.3
23	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3
24	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3
25	4.2	3.4	3.0	2.8	2.6	2.5	2.4	2.3	2.3	2.2
26	4.2	3.4	3.0	2.7	2.6	2.5	2.4	2.3	2.3	2.2
27	4.2	3.4	3.0	2.7	2.6	2.5	2.4	2.3	2.3	2.2
28	4.2	3.3	2.9	2.7	2.6	2.4	2.4	2.3	2.2	2.2
29	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2
30	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2

Table 1: F-kvantiler $F_{0.05}(f_1, f_2)$ avrundade till en decimals noggrannhet