STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

28 november 2019 9–14

Examinator: Ola Hössjer, tel. 070/672 12 18, ola@math.su.se
Återlämning: Meddelas via kurshemsida och webbaserat kursforum.
Tillåtna hjälpmedel: Miniräknare och formelsamling delas ut vid tentamenstillfället. Tabell över F-kvantiler återfinns nedan. Det gäller även att $\chi^2_{0.05}(1) \approx 3.8$.
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 1

Vattnet i en viss region innehåller två skadliga bakterietyper A och B. Halten av B-bakterier (enhet: mg/l) varierar mellan områden i regionen, medan halten A-bakterier kan anses vara konstant. Biologen Lisa har en utrustning som medger mätning av den totala bakteriehalten för vattenprover som genomgått ett visst reningsfilter. Detta filter lyckas eliminera andelen β av alla B-bakterier, medan alla typ A-bakterier passerar igenom filtret. Lisa analyserade vattenprover vid 16 olika fältstationer inom regionen och ställde upp en enkel linjär regressionsmodell

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, 16, \tag{1}$$

för sambandet mellan bakteriehalten Y_i för provet vid station $i = 1, \ldots, 16$ *efter* rening och andelen x_i av typ B-bakterier vid station i *före* rening ($x_1, \ldots, x_{16}$ hade tidigare bestämts med hjälp av en noggrannare mätutrustning). Således svarar α mot halten av typ A-bakterier i hela regionen. Lisa

antog vidare att ε_i är oberoende och $N(0, \sigma^2)$-fördelade feltermer. Resultatet från de 16 mätningarna sammanfattades i form av följande fem summor;

$$\begin{aligned} \textstyle\sum_{i=1}^{16} x_i &= 168.0, \\ \textstyle\sum_{i=1}^{16} (x_i - \bar{x})^2 &= 45.0, \\ \textstyle\sum_{i=1}^{16} Y_i &= 312.0, \\ \textstyle\sum_{i=1}^{16} Y_i(x_i - \bar{x}) &= 9.0, \\ \textstyle\sum_{i=1}^{16} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 &= 19.0, \end{aligned}$$

där $\hat{\alpha}$ och $\hat{\beta}$ är minsta kvadratskattningarna av halten A-bakterier respektive förmågan hos filtret att eliminera B-bakterier.

a) Beräkna $\hat{\alpha}$ och $\hat{\beta}$. (Ledning: Analysera först en centrerad regressionsmodell med intercept α_c, där de förklarande variablerna ändrats från x_i i den icke-centrerade parametriseringen (1), till $x_i - \bar{x}$.) (3 p)

b) Bestäm medelfelet $d = \sqrt{\widehat{\mathrm{Var}}(\hat{\alpha})}$ för $\hat{\alpha}$. (Ledning: Börja med att bestämma $\mathrm{Var}(\hat{\alpha})$.) (4 p)

c) Ange ett 95% konfidensintervall för α. (3 p)

Uppgift 2

Vid en stålindustri framställs en legering som består av tre metaller. Man vill ta reda på hur känslig legeringens hårdhet är för små variationer kring nuvarande värden av de tre ingående metallernas koncentration. Totalt genomförs $N = 20$ experiment $i = 1, \ldots, 20$ där hårdheten Y_i och koncentrationerna x_{1i}, x_{2i}, x_{3i} av de tre metallerna registreras. Man antar att metallerna påverkar legeringens hårdhet oberoende av varandra. Därför bortses från samspel mellan metallernas inverkan på legeringens hårdhet och olika linjära regressionsmodeller med ingen, en, två eller tre metallkoncentrationer som förklarande variabler jämförs. Dessutom är försöket upplagt så att effekterna av de olika förklarande variablerna är ortogonala, det vill säga

$$\sum_{i=1}^{20} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) = 0$$

för alla $1 \leq j < k \leq 3$, med $\bar{x}_j = \sum_i x_{ji}/20$. För den *fullständiga modellen* med alla tre metaller får man följande variansanalystabell:

Variationskälla	Kvs
Metall 1	3.1
Metall 2	5.1
Metall 3	8.9
Residual	25.0
Totalt	42.1

a) Genomför första steget i framåtinkludering (Forward Selection, FS). Undersök alltså om någon förklarande variabel ska tas med. Signifikansnivån väljs till 5%. (Ledning: På grund av ortogonaliteten mellan de förklarande variablerna fås kvadratsumman för avvikelsen mellan en grund- och en hypotesmodell som summan av kvadratsummorna (i tabellen ovan) för de metaller som ingår i grundmodellen men inte i hypotesmodellen. För varje delmodell av den fullständiga modellen ovan så inkluderas de metaller som inte ingår i delmodellen i variationskällan Residual för delmodellen.) (5 p)

b) Stannar FS-schemat efter a)? Motivera ditt svar. (5 p)

Uppgift 3

Vid ett medicinskt laboratorium mäter man diametern hos röda blodkroppar med hjälp av ett mikroskop (enhet: μm). Man vill uppskatta hur mycket blodkropparna hos en individ varierar i storlek. Vid ett tillfälle valde man ut 6 blodkroppar från en patient och genomförde 4 mätningar på varje blodkropp. De uppmätta diametrarna $Y_{i1}, \ldots, Y_{i4}$ för blodkropp $i = 1, \ldots, 6$ antas följa en ensidig variansanalysmodell

$$Y_{ij} = \mu + \delta_i + \varepsilon_{ij}$$

av typ II, med väntevärde μ, oberoende effekter $\delta_i \sim N(0, \sigma_\delta^2)$ av blodkropparna, samt oberoende mätfel $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. Resultatet sammanfattas i följande tabell, i form av stickprovsmedelvärden $\bar{Y}_{i\cdot}$ och stickprovsvarianser s_i^2 för mätningarna på respektive blodkropp:

Blodkropp i	$\bar{Y}_{i\cdot}$	s_i^2
1	7.5	0.020
2	7.1	0.015
3	8.2	0.025
4	7.4	0.018
5	7.9	0.022
6	7.5	0.014

a) Skatta de två varianskomponenterna σ_δ^2 och σ_ε^2. (5 p)
b) Ange ett 95% konfidensintervall för μ. (5 p)

Uppgift 4

Konditorn Pelle bakar en viss sorts tårta. För att öka försäljningen av denna tårttyp prövar han att variera mängden av tre ingående ingredienser - grädde (G), sylt (S) och maräng (M). Pelle genomför ett 2^3-försök utan replikat, där alla tre faktorerna varieras på en låg (-) och en hög (+) nivå, svarande mot en något lägre eller högre mängd av respektive ingrediens jämfört med

den tårttyp som för närvarande säljs. Pelle låter 8 kunder provsmaka tårtor med olika mängder av de tre ingredienserna. Han antar en additiv modell

$$Y_{ijk} = \mu + \bar{G} \cdot i + \bar{S} \cdot j + \bar{M} \cdot k + \varepsilon_{ijk}, \quad i, j, k \in \{-, +\} \tag{2}$$

för nöjdheten Y_{ijk} (mätt på en kontinuerlig skala mellan 0 och 10) hos den kund som prövar en tårta där G, S och M är på nivå i, j respektive k. Här anger μ genomsnittsnivån för hela försöket, medan $\bar{G}$, $\bar{S}$, $\bar{M}$ anger inflytandet hos de tre faktorerna. Feltermerna $\varepsilon_{ijk} \sim N(0, \sigma^2)$ antas oberoende. Resultatet från de 8 försökspunktera i, j, k framgår av följande tabell:

i,j,k	Y_{ijk}	i,j,k	Y_{ijk}
-,-,-	5.6	-,-,+	5.3
+,-,-	6.2	+,-,+	5.5
-,+,-	6.5	-,+,+	5.7
+,+,-	6.8	+,+,+	7.1

a) Beräkna minsta kvadrat-skattningarna $\hat{G}$, $\hat{S}$ och $\hat{M}$ av de tre faktorernas inflytande. (3 p)

b) Motivera att de tre skattningarna i a) är väntevärdesriktiga, normalfördelade och sinsemellan oberoende, med samma varians $\sigma^2/8$. (3 p)

c) Genomför ett F-test på nivån 5% för att avgöra om små variationer av mängden av de tre ingredienserna har någon signifikant inverkan på kundnöjdheten. Dvs testa grundmodellen (2) mot en hypotesmodell H_0 : $\bar{G} = \bar{S} = \bar{M} = 0$. (Ledning: Börja med att skatta σ^2. Du kan utnyttja att $\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2 = 3.0687$, samt att Kvs(Faktor $= \theta$) $= 8\hat{\theta}^2$ för en variationskälla som består av en faktor $\theta \in \{G, S, M\}$, där $\hat{\theta}$ anger motsvarande skattning från a). För en variationskälla bestående av *flera* faktorer kan du utnyttja att faktorerna är ortogonala i ett 2^3-försök.) (4 p)

Uppgift 5

Anta att vi har sex observationer

$$Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, \ldots, 6, \tag{3}$$

av en responsvariabel Y_i och två förklarande variabler/kovariater x_{1i} och x_{2i}, för en linjär modell utan intercept, där β_1 och β_2 är effektparametrarna för de två kovariaterna och där feltermerna $\varepsilon_i \sim N(0, \sigma^2)$ antas oberoende. Värdena på de förklarande variablerna för de sex försökspunkterna återfinns i följande tabell:

i	x_{1i}	x_{2i}
1	0	0
2	0	0
3	1	1
4	-1	1
5	-1	-1
6	1	-1

Med hjälp av detta dataset vill man prediktera värdet på en ny observation

$$Y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon = \mu + \varepsilon,$$

där designpunkten (x_1, x_2) är känd, och där $\varepsilon \sim N(0, \sigma^2)$ är oberoende av feltermerna ε_i i (3).

a) Härled den tvådimensionella normalfördelningen för minsta kvadrat-skattningen $(\hat{\beta}_1, \hat{\beta}_2)^T$ av $(\beta_1, \beta_2)^T$ baserad på modellen i (3). (2 p)

b) Responsvariabeln för den nya observationen kan predikteras med $\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Använd a) för att bestämma medelkvadratfelet

$$\text{MSEP} = E\left[(Y - \hat{Y})^2\right] = E\left[(Y - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2)^2\right]$$

för prediktionen av den nya observationen. (Ledning: Ditt svar kommer att bero på x_1, x_2 och σ^2. Du kan utnyttja att prediktionsfelet $Y - \hat{Y}$ kan uttryckas med hjälp av ε och skattningsfelen $\hat{\beta}_1 - \beta_1$ och $\hat{\beta}_2 - \beta_1$ för de två effektparametrarna.) (3 p)

c) Anta att vi istället anpassar datamaterialet med sex observationer till en modell där endast den första kovariaten finns med, dvs

$$Y_i = \beta_1 x_{1i} + \epsilon_i, \quad i = 1, \ldots, 6, \tag{4}$$

där $\epsilon_i = \varepsilon_i + \beta_2 x_{2i}$ ses som feltermer. Visa att man i denna modell får samma minsta kvadrat-skattning $\hat{\beta}_1$ av β_1 som i a), och alltså en prediktor $\hat{\beta}_1 x_1$ av den nya observationen Y. Använd sedan detta till att bestämma medelkvadratfelet

$$\text{MSEP}_1 = E\left[(Y - \hat{\beta}_1 x_1)^2\right]$$

för prediktion av Y med hjälp av modellen (4), som endast baseras på kovariat 1. För vilka värden på (x_1, x_2) och $(\beta_1, \beta_2, \sigma^2)$ får man lägre prediktionsfel genom att använda den mindre (och felaktiga) modellen (4) jämfört med den korrekta modellen (3)? (5 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