STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

23 oktober 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

En statistiskintresserad mäklare, som förmedlar lägenheter i en medelstor svensk kommun, vill kunna prediktera lägenhetspriser åt sina kunder. För åstadkomma detta tar han reda på lägenhetspriset Y_i (enhet: Mkr) och

boarean x_i (enhet: m^2) för 25 slumpmässigt utvalda lägenheter som det senaste året sålts in kommunen. Han ansätter en linjär regressionsmodell

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 25, \tag{1}$$

för sambandet mellan x_i och Y_i, där $\bar{x} = \sum_{i=1}^{25} x_i/25$ är det genomsnittliga priset för de undersökta lägenheterna och ε_i är oberoende och $N(0, \sigma^2)$-fördelade feltermer. Resultatet av mäklarens undersökning kan sammanfattas enligt följande:

$$\begin{aligned} \textstyle\sum_{i=1}^{25} x_i &= 2000, \\ \textstyle\sum_{i=1}^{25} (x_i - \bar{x})^2 &= 2800, \\ \textstyle\sum_{i=1}^{25} Y_i &= 48.0, \\ \textstyle\sum_{i=1}^{25} Y_i(x_i - \bar{x}) &= 83.0, \\ \text{Kvs(Residual)} &= 0.30. \end{aligned}$$

a) Beräkna minsta kvadrat-skattningarna $\hat{\tilde{\alpha}}$ och $\hat{\beta}$ av $\tilde{\alpha}$ och β. (3 p)

b) Lisa är med på en visning för en lägenhet på 90 kvadratmeter och hon ber mäklaren prediktera dess pris Y efter budgivning. Mäklaren börjar med att skatta det förväntade lägenhetspriset $E(Y) = \mu = \tilde{\alpha} + (90 - \bar{x})\beta$ med $\hat{\mu} = \hat{\tilde{\alpha}} + (90 - \bar{x})\hat{\beta}$, där $\hat{\mu}$ också kan ses som en prediktion av Y. Ange ett uttryck för prediktionsfelsvariansen $\mathrm{Var}(Y - \hat{\mu})$, genom att utnyttja att Y och $\hat{\mu}$ är oberoende. (3 p)

c) Ange ett 95% prediktionsintervall för det pris Y som Lisa måste betala om hon ger sig in i budgivningen och sedan köper lägenheten. (Ledning: Börja med att beräkna $\hat{\mu}$ och en skattning av σ.) (4 p)

Uppgift 2

För att förbättra sin modell från uppgift 1 beslutar sig mäklaren för att ta med en till förklarande variabel x_2. Den är binär, med värdena 0 och 1 för lägenheter som inte har respektive har en balkong. Detta ger upphov till den multipla linjära regressionsmodellen

$$Y_i = \tilde{\alpha} + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \varepsilon_i, \quad i = 1, \ldots, 25, \tag{2}$$

där Y_i är lägenhetspriset, $x_{1i} = x_i$ är boarean för lägenhet i från uppgift 1, x_{2i} anger huruvida lägenhet i har en balkong, medan ε_i är oberoende och $N(0, \sigma^2)$-fördelade feltermer. (Notera att $\tilde{\alpha}$ skattas som i uppgift 1. Däremot skattas inte β_1 och σ^2 i uppgift 2 som β och σ^2 i uppgift 1, eftersom vi tagit med en till förklarande variabel i uppgift 2.)

a) Formulera (2) på matrisform, där du speciellt definierar responsvektorn $\boldsymbol{Y}$ och x-delen $\boldsymbol{X}$ av designmatrisen. (2 p)

b) Mäklaren noterar att 15 av lägenheterna har balkong ($\bar{x}_2 = 0.6 = 15/25$) och räknar sedan ut

$$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 2800 & 60 \\ 60 & 6.0 \end{pmatrix}.$$

För att ta reda på hur pass kolineära de två förklarande variablerna är vill han beräkna variansinflationsfaktorn VIF för skattningen av β_1. Utför den beräkningen åt honom. (3 p)

c) En anpassning av modellen (2) ger samma skattning $\hat{\tilde{\alpha}}$ av $\tilde{\alpha}$ som i uppgift 1, samt $\hat{\beta}_1 = 0.02$ och $\hat{\beta}_2 = 0.2$. Beräkna förklaringsgraden R^2 för modellen. (Ledning: För att beräkna Kvs(Total) behöver du använda information från uppgift 1, medan informationen från uppgift 2b-2c) kan användas för att beräkna Kvs(Regression).) (5 p)

Uppgift 3

En biostatistiker vill undersöka om kost- och livsstilsvanor har en gemensam påverkan på Body Mass Index (BMI). Hon genomför en tresidig variansanalys typ I, där förutom kost (två nivåer) och livsstil (tre nivåer) även en tredje genetisk faktor med tre nivåer ingår. Totalt bestäms BMI för 36 individer, med två personer för varje nivåkombination av de tre faktorerna.

a) Formulera variansanalysmodellen matematiskt, där de tre huvudeffekterna ingår samt andra ordningens samspel mellan kost- och livsstilsfaktorerna. Ange också linjära restriktioner för de ingående parametrarna (för att undvika överparametrisering). (3 p)

b) En variansanalystabell från försöket har följande utseende:

Variationskälla	Kvs	f
Kost	6.0	
Livsstil	12.0	
Genetisk	5.0	
Samspel Kost och Livsstil	10.0	
Residual	28.0	
Totalt	61.0	

Börja med att fylla i antalet frihetsgrader f i tabellens tredje kolumn. Testa sedan på nivån 5% om det finns något signifikant samspel mellan faktorerna Kost och Livsstil. (4 p)

c) Testa på nivån 5% om den genetiska faktorn har en signifikant påverkan på BMI. (3 p)

Uppgift 4

Erik vill bestämma vikterna θ_1 och θ_2 (enhet: kg) av två silverljusstakar som han hittat på sin farmors vind. Till sin hjälp har han en balansvåg med två skålar A och B, som han också hittat på samma vind. Vid varje mätning ger vågen ett utslag som är differensen mellan vikterna i skål A och B, plus ett normalfördelat mätfel med väntevärde 0 och varians σ^2. Vidare antas mätfelen vara oberoende mellan olika mätningar. Erik utför en mätserie enligt följande:

Mätning	Skål A	Skål B	Vågens utslag
1	1 och 2	Inga	5.1
2	Inga	1 och 2	-5.2
3	1	2	1.2
4	2	1	-0.9

a) Formulera Eriks försök som en allmän linjär modell med responsvektor $\boldsymbol{Y}$, designmatris $\boldsymbol{A}$, parametervektor $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ och feltermsvektor $\boldsymbol{\varepsilon}$. (2 p)

b) Bestäm minsta kvadrat-skattningen $\hat{\boldsymbol{\theta}}$ av $\boldsymbol{\theta}$ samt kovariansmatrisen $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ för denna skattning uttryckt i feltermsvariansen. (3 p)

c) Erik ber sin lillebror Sten upprepa försöket, men Sten lägger endast en ljusstake på vågen vid varje mätning. Hur många mätningar måste Sten i så fall totalt utföra för att kunna skatta dels θ_1 och dels θ_2 med samma noggrannhet (dvs samma varians) som för Eriks mätningar? (Inga långa räkningar krävs, men du måste motivera ditt svar.) (1 p)

d) Bestäm konfidenscirkeln (alltså en cirkelformad konfidensregion) för $\boldsymbol{\theta}$ med konfidensgrad 95%. Du får utan motivering utnyttja att Kvs(Residual) = 0.05. (4 p)

Uppgift 5

Vid en industri undersökte man hur utbytet från en kemisk reaktion varierade då koncentrationen av de 5 ingående substanserna A, B, C, D och E ändrades. Man genomförde ett fraktionellt 2^{5-2}-försök med 8 mätningar, där varje substans koncentration antingen låg på en högre eller lägre nivå än den nivå som då användes i fabriken. Man ställde alltså upp modellen

$$Y_{ijklm} = \mu + \bar{A} \cdot i + \bar{B} \cdot j + \bar{C} \cdot k + \bar{D} \cdot l + \bar{E} \cdot m + \varepsilon_{ijklm}, \qquad (3)$$

för utbytet av reaktionen då de fem faktorerna var på nivåerna $i, j, k, l, m \in \{-, +\}$. Här svarar -, den låga nivån, mot -1, medan +, den höga nivån, svarar mot +1. Vidare antas att $\varepsilon_{ijklm} \sim N(0, \sigma^2)$ är oberoende och normalfördelade feltermer. Resultatet av undersökningen framgår av följande tabell:

A	B	C	D	E	Y_{ijklm}
-	-	-	-	+	11.0
+	+	-	-	+	16.0
+	-	+	-	-	18.0
-	+	+	-	-	14.0
+	-	-	+	-	18.0
-	+	-	+	-	12.0
-	-	+	+	+	21.0
+	+	+	+	+	26.0

a) Beräkna minsta kvadrat-skattningarna $\hat{A}$, $\hat{B}$, $\hat{C}$ $\hat{D}$ och $\hat{E}$ av de fem effektparametrarna $\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{E}$. (3 p)
b) Man kan visa att Mkvs(Residual) = 0.5. Använd denna information för att ge ett 95% konfidensintervall för den totala förväntade höjningen av utbytet, $\bar{A} + \bar{B} + \bar{C} + \bar{D} + \bar{E}$, om koncentrationen av alla fem substanser ändras från normalläget till den höga nivån. (Ledning: Använd dig av en lämplig linjärkombination av parametervektorn.) (3 p)
c) Bestäm vilka effekter i en fullständig modell (som förutom enheten och huvudeffekterna i (3) också har samspel av alla ordningar) som är kopplade till enheten. Bestäm sedan vilka andra ordningens samspel som kan läggas till i (3). (Ledning: Dessa samspel får inte vara kopplade till enheten eller till en huvudeffekt. Om flera samspel av ordning två är kopplade till varandra ska du endast ta med ett av dem.) (4 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