STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

23 oktober 2023 14–19

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

Ett läkemedelsföretag har tagit fram en astmamedicin och vill i en förstudie med 10 patienter undersöka om medicinen fungerar så pass väl att man senare kan gå vidare och genomföra större studier. För varje patient $i = 1, \ldots, 10$ i förstudien, som tidigare samma dag tagit x_i mg av medicinen, uppmättes utandningsvolymen Y_i liter under en sekund. Resultatet framgår av följande tabell:

Dos i mg	Utandningsvolum
1	5.0, 5.3
2	5.2, 5.7
3	5.0, 5.8
4	5.3, 6.0
5	5.6, 5.9

Man ansatte en linjär regressionsmodell

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i = \mu(x_i) + \varepsilon_i, \quad i = 1, \ldots, 10,$$

där $\mu(x) = \tilde{\alpha} + \beta(x - \bar{x})$ är den förväntade utandningsvolumen för en patient som tagit x mg av medicinen, $\bar{x} = \sum_{i=1}^{10} x_i/10$ är den genomsnittliga dosen för alla tio patienter, och ε_i är oberoende och normalförelarde feltermer med väntevärde 0 och standardavvikelse σ.

a) Beräkna en skattning $\hat{\mu}(6) = \hat{\tilde{\alpha}} + \hat{\beta}(6 - \bar{x})$ av den förväntade utandningsvolumen hos en patient som tar 6 mg av medicinen. (Ledning: $\sum_{i=1}^{10} y_i = 54.8$ och $\sum_{i=1}^{10} y_i(x_i - \bar{x}) = 2.8$.) (3 p)

b) Beräkna en skattning av σ^2, utgående från följande (ofullständiga) variansanalystabell:

Variationskälla	Kvs
Regression	0.392
Residual	
Total	1.216

(3 p)

c) Ange ett 95% prediktionsintervall för utandningvolymen hos en patient som tar 6 mg av medicinen. (Ledning: Prediktionsfelsvariansen är $\sigma^2 + \mathrm{Var}(\hat{\tilde{\alpha}}) + (6 - \bar{x})^2 \mathrm{Var}(\hat{\beta})$, samt $t_{p/2}(f) = \sqrt{F_p(1, f)}$, där F-kvantilen återfinns i tabellen nedan.) (4 p)

Uppgift 2

I en epidemiologisk studie ville man undersöka om en viss ämnesomsättningssjukdom endast orsakades av livsstilsfaktorer, eller om det också fanns ärftliga komponenter. Man uppmätte koncentrationen Y_i av en viss hormontyp i blodet som påverkar ämnesomsättningen, för 30 olika patienter ($i = 1, \ldots, 30$). För varje person uppmättes också tre livsstilsfaktorer (matvanor, stressnivå, infektionskänslighet) x_{1i}, x_{2i}, x_{3i}, samt vilka varianter x_{4i}, x_{5i} av två olika gener som ärvts ned.

Man ansatte en multipel linjär regressionsmodell (grundmodellen) med parametervektor $\boldsymbol{\theta} = (\alpha, \beta_1, \ldots, \beta_5)^T$, där α är ett icke-centrerat intercept, β_j anger effekten av förklarande variabel j, och där feltermerna antas oberoende och normalfördelade med väntevärde 0 och varians σ^2.

a) Formulera hypotesmodellen att bara livsstilsfaktorer påverkar ämnesomsättningen, enligt $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\lambda}$, för lämpligt vald matris $\boldsymbol{B}$ och kolumnvektor $\boldsymbol{\lambda}$. (Ledning: $\boldsymbol{B}$ har dimensionen 6×4 och $\boldsymbol{\lambda}$ har 4 komponenter.) (2 p)

b) Definiera förklaringsgraderna R_0^2 och R_1^2 för grund- och hypotesmodellen med hjälp av Y_i, $\hat{\mu}_i$ och $\hat{\tilde{\mu}}_i$ för alla patienter $i = 1, \ldots, 30$, samt $\bar{Y} = \sum_{i=1}^{30} Y_i/30$. Här anger $\hat{\mu}_i$ och $\hat{\tilde{\mu}}_i$ skattningar av $\mu_i = E(Y_i)$ från grund- respektive hypotesmodellen. (2 p)

c) Resultatet av studien gav en förklaringsgrad $R_0^2 = 0.751$ för grundmodellen, och $R_1^2 = 0.682$ för hypotesmodellen. Testa på nivån 5% om genetiska

faktorer har en signifikant inverkan på ämnesomsättningen. (Ledning: Börja med att titta på $R_0^2 - R_1^2$ och $1 - R_0^2$ och bilda sedan deras kvot.) (3 p)

d) Ange en väntevärdesriktig skattning av σ^2, genom att utnyttja värdet på R_0^2 och extrainformationen Kvs(Total) = 41.2. (3 p)

Uppgift 3

En mäklarfirma ville undersöka hur mycket priset av en viss typ av husobjekt berodde på dess läge. Eftersom det var fråga om en exlusiv och ovanlig hustyp hade man endast försäljningspris från fem tidigare sålda objekt. Man ansatte en multipel linjär regressionsmodell

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

för priset på hus $i = 1, \ldots, 5$, där de två förklarande variablerna samvarierade enligt följande tabell:

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

Här anger x_{1i} husets läge svarande mot glesbygd (-1), förort (0) och tätort (1), medan x_{2i} anger graden av efterfrågan när huset såldes, svarande mot lågsäsong (-1), normalsäsong (0) och högsäsong (1). Parametern av primärt intresse är β_1, medan β_2 finns med för att skapa en mer tillförlitlig modell, till priset av högre varians för parameterskattningarna. För att inte överanpassa modellen tog man inte med fler än två förklarande variabler. Feltermerna $\varepsilon_i \sim N(0, \sigma^2)$ antogs vara oberoende.

a) Ange ett uttryck för $\mathrm{Var}(\hat{\beta}_1)$, förenklat så långt som möjligt. (Ledning: Börja med att bestämma designmatrisen $\boldsymbol{A}$ och därefter $\boldsymbol{X}^T\boldsymbol{X}$, där $\boldsymbol{X}$ är den del av designmatrisen som härrör från de förklarande variablerna. Matrisinversionsformeln

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

kan också vara användbar.) (4 p)

b) Hur hade uttrycket för $\mathrm{Var}(\hat{\beta}_1)$ sett ut om β_2 varit känd? (Ledning: Om $\beta_2 x_{2i}$ är känd kan detta uttryck subtraheras bort från Y_i.) (3 p)

c) Bestäm variansinflationsfaktorn (VIF) för skattningen av β_1, förenklat så långt som möjligt. (3 p)

Uppgift 4

Träden i ett visst område har utsatts för olika grad av svampangrepp, som också varierar mellan säsonger beroende på vädret. Vid ett lantbruksuniversitet ville man studera detta närmare och mätte graden av svampangrepp för tre slumpmässigt valda träd, som analyserades under tre säsonger, slumpmässigt valda från en tioårsperiod. Men ansatte en tvåsidig variansanalysmodell för graden

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad 1 \leq i, j, k \leq 3,$$

av svampangrepp på barkprov k, taget från träd i under säsong j. Eftersom de tre träden och säsongerna valdes ut som stickprov antog man att dessa faktorer var slumpmässiga (typ II), med möjligt samspel. Det innebär att alla α_i, β_j och γ_{ij} är oberoende stokastiska variabler med varians σ_α^2, σ_β^2 respektive σ_γ^2. Man var primärt intresserad av att skatta samspelsvariansen σ_γ^2, för att ta reda på om vissa träd var mer känsliga för vädervariation än andra. Feltermerna ε_{ijk} anger variationen mellan olika barkbitar från samma träd. De antogs vara oberoende och normalfördelade med väntevärde 0 och varians σ^2.

Resultatet av försöket framgår av följande två tabeller, där $\bar{Y}_{ij\cdot}$ och s_{ij}^2 anger stickprovsmedelvärdena och stickprovsvarianserna för de tre barkproven som togs från träd i under säsong j (dessutom anges radmedelvärden $\bar{Y}_{i\cdot\cdot}$, kolumnmedelvärden $\bar{Y}_{\cdot j\cdot}$ och totalmedelvärde $\bar{Y}_{\cdots}$ av alla stickprovsmedelvärden):

Stickprovsmedelvärden $\bar{Y}_{ij\cdot}$:

	$j=1$	$j=2$	$j=3$	$\bar{Y}_{i\cdot\cdot}$
$i=1$	1.8	3.1	5.0	3.3
$i=2$	2.6	3.5	9.2	5.1
$i=3$	0.7	1.8	2.9	1.8
$\bar{Y}_{\cdot j\cdot}$	1.7	2.8	5.7	$\bar{Y}_{\cdots} = 3.4$

Stickprovsvarianser s_{ij}^2:

	$j=1$	$j=2$	$j=3$
$i=1$	0.32	0.44	0.64
$i=2$	0.52	0.56	0.84
$i=3$	0.28	0.44	0.64

a) Beräkna en skattning av variansen σ^2 inom celler, genom att vikta ihop värdena i den högra tabellen. (4 p)

b) Beräkna medelkvadratsumman för variationskällan samspel, och testa sedan om samspelsvariansen är signifikant skild från 0 på nivån 5%. (Ledning: Börja med att beräkna skattningar $\hat{\gamma}_{ij}$ av alla γ_{ij}, genom att utgå från värden på $\bar{Y}_{ij\cdot}$, $\bar{Y}_{i\cdot\cdot}$, $\bar{Y}_{\cdot j\cdot}$ och $\bar{Y}_{\cdots}$ i den vänstra tabellen. Vid test av samspel har du sedan nytta av uträkningen in a).) (3 p)

c) Skatta samspelsvariansen σ_γ^2 mellan träd och säsong. (Ledning: Du kan ha nytta av a)-b), samt formeln för $E\,[\text{Mkvs(Samspel)}]$ vid tvåsidig variansanalys av typ II.) (3 p)

Uppgift 5

En ARMA$(1,1)$-process $\{X_t\}$ ges av

$$X_t - \phi X_{t-1} = \varepsilon_t - \theta \varepsilon_{t-1} \tag{1}$$

för $t = \ldots, -1, 0, 1, \ldots$, där $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ är oberoende feltermer.

a) För vilka värden på parametrarna ϕ och θ är processen stationär? (Inga långa uträkningar krävs.) (2 p)

b) Anta att ϕ och θ valts enligt a) så att $\{X_t\}$ är en stationär process. Härled uttryck för kovariansfunktionen $\gamma_k = \mathrm{Cov}(X_t, X_{t+k})$ för $k = 0$ och $k = 1$. (Ledning: Du kan lösa b) utan att först ha löst a). Utnyttja $\mathrm{Cov}(X_t, \varepsilon_{t+1}) = 0$. Använd sedan (1) för att ge ett uttryck för $\mathrm{Cov}(X_t, \varepsilon_t)$, innan du använder (1) igen för att beräkna γ_0 och γ_1.) (5 p)

c) Givet ϕ finns det två värden på θ sådana att $\gamma_1 = 0$. Endast för det ena av dessa två värden på θ gäller $X_t = \varepsilon_t$. Vilket? Motivera ditt svar. (Ledning: Använd b) och (1), omskrivet med bakåtoperatorn B, det vill säga den operator som uppfyller $BX_t = X_{t-1}$ och $B\varepsilon_t = \varepsilon_{t-1}$.) (3 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