STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

11 december 2024 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 visst land har genomgått en snabb ekonomisk utveckling. BNP-ökningen i procent de senaste 9 åren framgår av följande tabell:

År	BNP-ökning	År	BNP-ökning
2015	2.1	2020	4.6
2016	2.8	2021	4.5
2017	3.2	2022	4.9
2018	3.0	2023	5.5
2019	3.9		

Forskare vid den statistiska centralbyrån i landet ansätter en enkel linjär regressionsmodell

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 9,$$

för BNP-ökningen år x_i, där $\varepsilon_1, \ldots, \varepsilon_9$ är oberoende och normalfördelade feltermer med väntevärde 0 och varians σ^2, och $\bar{x} = \sum_{i=1}^{9} x_i/9 = 2019$.

Man är framför allt intresserade av att uppskatta den årliga ökningstakten β av BNP-ökningen.

a) Beräkna minsta kvadratskattningen $\hat{\beta}$ av β. (Ledning: $\sum_{i=1}^{9}(x_i - \bar{x})y_i = 24.1$, där y_i är de observerade värdena av Y_i, från tabellen ovan.) (3 p)

b) Från en variansanalystabell med variationskällorna Regression och Residual, avläser man att Kvs(Total) = 10.12 och Kvs(Regression) = 9.68. Bestäm härrur en väntevärdesriktig skattning av feltermsvariansen σ^2. (Ledning: Kvs(Total) är summan av kvadratsummorna för de två ovannämnda variationskällorna.) (3 p)

c) Beräkna ett 95% konfidensintervall för β. Är ökningen av landets BNP-tillväxt signifikant på nivån 5%? (Ledning: Du kan utnyttja att $t_{0.025}(f) = \sqrt{F_{0.05}(1, f)}$ för lämpligt antal frihetsgrader f, där F-kvantilen fås ur den bifogade tabellen.) (4 p)

Uppgift 2

Två företag har utvecklat var sin maskin för vattenrening, som båda tar bort en stor del av två typer av bakterier från vattnet. Hälsovårdsmyndigheten i ett land ville underöka om maskinerna är likvärdiga, och testade därför de två maskinerna på 6 olika vattenprover. Man ansatte därför den multipla linjära regressionsmodellen

$$Y_{ik} = \alpha + \beta_{i1}x_{1k} + \beta_{i2}x_{2k} + \varepsilon_{ik}, \quad i = 1, 2, k = 1, \ldots, 6 \tag{1}$$

för den totala bakteriehalten (enhet: antal bakteriekolonier mer 100 ml) i vattnet efter rening av prov nummer k med maskin i. Halten av respektive bakterietyp före rening var x_{1k} och x_{2k} i prov k, medan $0 \leq \beta_{ij} \leq 1$ anger hur stor andel av bakterierna av typ $j = 1, 2$ som *inte* elimineras med maskin i. Vidare anger α halten av övriga typer av bakterier (före och efter rening), som ingen av maskinerna lyckas filtrera bort, medan feltermerna $\varepsilon_{ik} \sim N(0, \sigma^2)$ är oberoende och normalfördelade.

a) Ange observationsvektor $\boldsymbol{Y}$, designmatris $\boldsymbol{A}$, parametervektor $\boldsymbol{\theta}$ och feltermsvektor $\boldsymbol{\varepsilon}$ för den linjära modellen

$$\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon}$$

som svarar mot (1). (Ledning: Antalet element Y_{ik} i kolumnvektorn $\boldsymbol{Y}$ är $N = 12$, och $\boldsymbol{\theta}$ är en kolumnvektor med fem element.) (3 p)

b) Betrakta (1) som vår grundmodell, som testas mot hypotesmodellen

$$H_0 : \beta_{11} = \beta_{21}, \beta_{12} = \beta_{22}$$

att de båda maskinerna är likvärdiga när det gäller att filtrera bort respektive bakterietyp. Denna hypotesmodell kan formularas som att parametervektorn i 2a) måste uppfylla $\boldsymbol{\theta} = \boldsymbol{B\lambda}$ för en viss matris $\boldsymbol{B}$ och kolumnvektor $\boldsymbol{\lambda}$. Ange $\boldsymbol{B}$ och $\boldsymbol{\lambda}$. (Ledning: Kolumnvektorn $\boldsymbol{\lambda}$ har tre element.) (3 p)

c) Nedan ges ett utdrag ur försökets variansanalystabell:

Variationskälla	Kvs
Avvikelse från H_0	8.00
Residual	8.14
Totalt	16.14

Använd denna information för att på signifikansnivån 5% testa om de båda vattenreningsmaskinerna är likvärdiga eller ej. (4 p)

Uppgift 3

Strålningsfysiker vid en kärnkraftsanläggning ville undersöka hur mycket strålningen från anrikat kärnbränsle (enhet: GBq/ton) varierade mellan olika bränslestavar, samt även uppskatta den genomsnittliga strålningsmängden μ. De genomförde 5 mätningar på var och en av 4 bränslestavar. Därefter ansatte de en ensidig variansanalysmodell typ II, enligt

$$Y_{ij} = \mu + \delta_i + \varepsilon_{ij}, \quad i = 1, 2, 3, 4,\ j = 1, 2, 3, 4, 5,$$

där Y_{ij} är den uppmätta stålningsmängden från den j:te mätningen av stav i, medan $\delta_i \sim N(0, \sigma_\delta^2)$ och $\varepsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ är oberoende stokastiska variabler, svarande mot slumpvariationen i strålningsmängd mellan stavar respektive variation mellan olika mätningar från samma stav. Resultaten av mätningarna framgår av följande tabell, där stickprovsmedelvärdena $\bar{Y}_{i\cdot}$ och stickprovsvarianserna s_i^2 är angivna för varje stav:

Stav i	$\bar{Y}_{i\cdot}$	s_i^2
1	80.0	0.10
2	82.0	0.14
3	78.5	0.08
4	79.5	0.12
Medel	80.0	0.11
Std	1.472	

De nedre två raderna anger även medelvärden och stickprovsstandardavvikelser från respektive kolumn.

a) Använd stickprovsvarianserna s_i^2 för att beräkna en skattning av mätfelsvariansen σ_ϵ^2. (3 p)

b) Beräkna en skattning av variansen σ_δ^2. (Ledning: Utnyttja 3a) och stickprovsstandardavvikelsen av radmedelvärdena $\bar{Y}_{i\cdot}$.) (4 p)

c) Beräkna ett 95% konfidensintervall för μ. (3 p)

Uppgift 4

En stationär MA(1)-process $\{X_t\}$ definieras genom

$$X_t = \varepsilon_t - \theta\varepsilon_{t-1} \tag{2}$$

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

a) Använd (2) för att beräkna kovariansfunktionen $\gamma_k = \mathrm{Cov}(X_t, X_{t+k})$ för $k = 0, 1, 2, \ldots$. (4 p)

b) Använda 4a) för att beräkna korrelationsfunktionen $\rho_k = \mathrm{Corr}(X_t, X_{t+k})$ för $k = 0, 1, 2, \ldots$. (3 p)

c) För varje korrelationsfunktion i 4b) finns det två värden på θ som ger denna korrelationsfunktion. Finns det någon skillnad mellan motsvarande två MA(1)-processer? (3 p)

Uppgift 5

Ett dataset med sex observationer Y_i av en responsvariabel och två förklarande variabler/kovariater x_{1i} och x_{2i}, finns sammanfattat i följande tabell:

i	x_{1i}	x_{2i}	Y_i
1	0	0	-0.1
2	0	0	0.1
3	1	1	2
4	-1	1	0
5	-1	-1	-2
6	1	-1	0

Man vill undersöka om en eller båda av kovariaterna kan förklara variationen i responsvariabeln. Därför införs den fulla linjära regressionsmodellen

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

där inget intercept men båda kovariaterna ingår, med effektparametrar β_1 respektive β_2, och där $\varepsilon_i \sim N(0, \sigma^2)$ är oberoende och normalfördelade feltermer. Man vill sedan jämföra anpassningen till data för (3) och tre andra delmodeller. För dessa tre delmodeller ingår ingen kovariat, bara kovariat 1 respektive bara kovariat 2.

a) Genomför första steget av framåtinkludering (Forward Selection, FS), genom att dels testa modellen utan någon kovariat mot modellen med kovariat 1, och dels testa modellen utan någon kovariat mot modellen med kovariat 2. (Ledning: Börja med att beräkna skattningar av β_j och σ^2 för modellen där bara kovariat j ingår, för $j = 1, 2$.) (4 p)

b) Genomför första steget av bakåteliminering (Backward Elimination, BE), genom att dels testa modellen med kovariat 1 mot (3), och dels testa modellen med kovariat 2 mot (3). (Ledning: Börja med att beräkna skattningar av β_1, β_2 och σ^2 för den fulla modellen (3). Använd sedan även skattningarna av β_1 och β_2 från 5a).) (4 p)

c) Vilken modell väljs enligt FS och BE? Förklara varför man får olika svar, och ange vilken modellvalsmetod som i detta fall är att föredra. (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