STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

9 december 2022 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

En forskargrupp ville undersöka om mindre träd av en viss art lättare drabbades av svampangrepp än större. Man delade upp trädbeståndet i sju grupper $x = 1, 2, 3, 4, 5, 6, 7$. Dessa värden var proportionella mot den genomsnittliga diametern hos trädstammarna i respektive grupp. Man uppmätte graden av svampangrepp Y_i för $N = 7$ träd, ett från varje grupp. Därefter ansattes en enkel linjär regressionsmodell

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 7,$$

med $x_i = i$, $\bar{x} = \sum_{i=1}^{7} x_i/7$ och feltermer $\varepsilon_i \sim N(0, \sigma^2)$ som antas oberoende. För att testa nollhypotesen $H_0 : \beta = 0$ användes följande utdrag ur en variansanalystabell:

Variationskälla	Kvs
Regression	9.20
Residual	5.75
Totalt	14.95

a) Räkna ut medelkvadratsummorna för de två variationskällorna och genomför därefter ett test av H_0 mot alternativhypotesen $H_1 : \beta \neq 0$ på signifikansnivån 5%. (6 p)

b) Beräkna minsta kvadrat-skattningen $\hat{\beta}$ av β, givet extrainformationen $\hat{\beta} < 0$. (Ledning: Börja med att beräkna $\hat{\beta}^2$ med hjälp av kvadratsumman för Regression i variansanalystabellen.) (4 p)

Uppgift 2

Vid en industri undersöktes huruvida utbytet av en kemisk reaktion påverkades av tre olika katalysatorer. Utbytet Y_i och koncentrationerna x_{1i}, x_{2i}, x_{3i} av de tre katalysatorerna registrerades vid $N = 10$ olika experiment ($i = 1, \ldots, 10$). De tre katalysatorerna antogs verka oberoende av varandra. Därför bortsåg man från samspel och jämförde olika linjära regressionsmodeller med ingen, en, två eller tre katalysatorkoncentrationer som förklarande variabler. Dessutom var försöket upplagt så att effekterna av de olika förklarande variablerna var ortogonala mot varandra, det vill säga

$$\sum_{i=1}^{10}(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) = 0$$

för alla $1 \leq j < k \leq 3$, med $\bar{x}_j = \sum_i x_{ji}/10$. För den *fullständiga modellen* med alla tre katalysatorer fick man följande variansanalystabell:

Variationskälla	Kvs
Katalysator 1	4.1
Katalysator 2	6.2
Katalysator 3	2.1
Residual	2.8
Totalt	15.2

a) Bestäm antal frihetsgrader för alla variationskällor i variansanalystabellen. (2 p)

b) Genomför första steget i framåtinkludering (Forward Selection, FS), det vill säga undersök om någon förklarande variabel ska tas med. Signifikansnivån väljs till 5%. (Ledning: På grund av ortogonaliteten mellan de förklarande variablerna fås kvadratsumman för regressionsdelen av en *delmodells* variansanalystabell (det vill säga F-kvotens täljare) genom att addera kvadratsummorna för de katalysatorer som ingår i delmodellen, medan resterande kvadratsummor från den fullständiga modellen tillhör variationskällan residual för delmodellen.) (5 p)

c) Stannar FS-schemat efter a)? Motivera ditt svar. (3 p)

Uppgift 3

I ett visst land vaccineras alla barn under sitt första levnadsår mot en allvarlig sjukdom. Hälften av dem får Vaccin 1, medan den andra hälften får Vaccin 2. De båda vaccinerna ger likvärdigt skydd mot sjukdomen, men man misstänker att de kan ge olika biverkningar i form av feber de närmsta dygnen efter vaccineringen. Ett läkemedelsbolag har nyligen utvecklat ett tredje vaccin (med samma sjukdomsskydd) som man hävdar ger lägre feber än de två andra vaccinerna. Smittskyddsinstitutet i landet genomför en mindre pilotstudie för att jämföra biverkningarna av de tre vaccinerna. Den består av 18 olika barn som slumpmässigt delas in i tre lika stora grupper, där barnen i varje grupp ges samma vaccin. Men registrerar sedan den maximala febern Y_{ij} för barn $j = 1, \ldots, 6$ som erhöll Vaccin $i = 1, 2, 3$. Resultatet framgår av följande tabell, där medelvärdet $\bar{Y}_{i\cdot}$ och stickprovsstandardavvikelsen s_i anges för respekive vaccin:

Vaccin	$\bar{Y}_{i\cdot}$	s_i
1	38.1	0.25
2	38.4	0.31
3	37.8	0.19

a) Anta att försöket beskrivs av en ensidig typ I variansanalysmodell, med $\mu_i = E(Y_{ij})$. Testa på nivån 5% nollhypotesen $\mu_1 = \mu_2 = \mu_3$, genom att först beräkna lämpliga medelkvadratsummor. (4 p)

b) Smittskyddsforskarna överväger att ersätta de två vaccinerna som för närvarande används med Vaccin 3. Beräkna en punktskattning $\hat{\Delta}$ av den förväntade minskningen $\Delta = (\mu_1 + \mu_2)/2 - \mu_3$ av feber i hela populationen om ett sådant vaccinbyte görs. (2 p)

c) Beräkna $\mathrm{Var}(\hat{\Delta})$ uttryckt i feltermsvariansen σ^2, och ange därefter medelfelet för $\hat{\Delta}$. (Ledning: Använd en skattning av σ^2 från den medelkvadratsumma i a) som baseras på stickprovsvarianserna, som ett led i att beräkna medelfelet för $\hat{\Delta}$.) (4 p)

Uppgift 4

Låt $\{Y_t\}$ vara en stationär AR(2)-process, det vill säga $X_t = Y_t - E(Y_t) = Y_t - \mu$ ges av

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t, \tag{1}$$

med oberoende feltermer $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, medan ϕ_1 och ϕ_2 är de två autoregressionsparametrarna.

a) Låt $\rho_k = \mathrm{Corr}(Y_t, Y_{t+k}) = \mathrm{Corr}(X_t, X_{t+k})$, $k = 0, 1, 2, \ldots$ beteckna autokorrelationsfunktionen. Visa att

$$\begin{aligned} \rho_1 &= \phi_1/(1-\phi_2), \\ \rho_2 &= \phi_2 + \phi_1^2/(1-\phi_2). \end{aligned}$$

(Ledning: Börja med att utnyttja (1) för att ge uttryck för $\gamma_1 = \mathrm{Cov}(X_t, X_{t+1})$ och $\gamma_2 = \mathrm{Cov}(X_t, X_{t+2})$ som innefattar $\gamma_0 = \mathrm{Var}(X_t)$. Uttryck därefter ρ_k med hjälp av γ_k och γ_0.) (5 p)

b) Anta att vi har data $y_1, \ldots, y_T$ från tidsserien i a). Definiera skattningen $\hat{\rho}_k$ av autokorrelationsfunktionen för $0 \leq k < T$. (2 p)

c) Anta att vi vill testa nollhypotesen

$$H_0 : \phi_1 = \phi_2 = 0.$$

Vi ser från a) att nollhypotesen medför $\rho_1 = \rho_2 = 0$. Konstruera ett test som innefattar $\hat{\rho}_1$ och $\hat{\rho}_2$ och som för stora värden på T approximativt har signifikansnivån 5%. (Ledning: Du får utan bevis använda att $\sqrt{T}\hat{\rho}_k \sim N(0,1)$ gäller approximativt under nollhypotesen då T är stor.) (3 p)

Uppgift 5

Anta att vi har en multipel linjär regressionsmodell, där varje observation

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, N,$$

kan skrivas som en summa av ett väntevärde μ_i och en felterm $\varepsilon_i \sim N(0, \sigma^2)$.

a) Visa hur μ_i beror linjärt av de m förklarande variablerna $x_{1i}, \ldots x_{mi}$ för observation i genom att införa lämpliga regressionsparametrar. (2 p)

b) Definera förklaringsgraden R^2 med hjälp av observationerna Y_i och minsta kvadrat-skattningarna $\hat{\mu}_i$ av alla μ_i. (3 p)

c) Korrelationskoefficienten mellan två vektorer $\boldsymbol{a} = (a_1, \ldots, a_N)^T$ och $\boldsymbol{b} = (b_1, \ldots, b_N)^T$ kan skrivas som

$$\mathrm{Corr}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(b_i - \bar{b})^2}},$$

där $\bar{a} = \sum_{i=1}^{N} a_i/N$ och $\bar{b} = \sum_{i=1}^{N} b_i/N$. Visa att

$$R^2 = \mathrm{Corr}(\boldsymbol{Y}, \hat{\boldsymbol{\mu}})^2,$$

där $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ och $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_N)^T$. (Ledning: Du kan utnyttja att $\sum_{i=1}^{N} Y_i/N = \sum_{i=1}^{N} \hat{\mu}_i/N = \bar{Y}$, samt att residualvektorn $\boldsymbol{Y} - \hat{\boldsymbol{\mu}}$ är ortogonal mot den skattade väntevärdesvektorn $\hat{\boldsymbol{\mu}}$.) (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