STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

17 april 2024 8–13

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

I ett kemiskt laboratorium uppmättes hållfastheten Y_i för $i = 1, \ldots, 12$ olika prover av en metallegering. Man lyckades inte helt frigöra beståndsdelarna i legeringen från kontamination, utan mätning i innehöll koncentrationen x_i av andra ämnen. För att korrigera för kontaminationens påverkan på hållfastheten satte man upp en enkel linjär regressionsmodell

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, 12, \tag{1}$$

för sambandet mellan hållfasthet och koncentrationen av andra ämnen. Här är ε_i oberoende och $N(0, \sigma^2)$-fördelade feltermer, α svarar mot hållfastheten för legeringen utan kontamination, medan lutningsparametern β anger hållfasthetens känslighet för kontamination. Resultatet från de 12 mätningarna sammanfattades i form av fem summor;

$$\begin{aligned}
\textstyle\sum_{i=1}^{12} x_i &= 20.5, \\
\textstyle\sum_{i=1}^{12} (x_i - \bar{x})^2 &= 13.2, \\
\textstyle\sum_{i=1}^{12} Y_i &= 31.4, \\
\textstyle\sum_{i=1}^{12} Y_i (x_i - \bar{x}) &= -4.1, \\
\textstyle\sum_{i=1}^{12} (Y_i - \bar{Y})^2 &= 3.3.
\end{aligned}$$

a) Beräkna minsta kvadrat-skattningen $\hat{\alpha}$ av det icke-centrerade interceptet α, det vill säga minsta kvadrat-skattningen av den okontaminerade metallegeringens hållfasthet. (Ledning: Skatta först parametrarna i en regressionsmodell med lutningsparameter β och centrerat intercept $\tilde{\alpha}$, där de förklarande variablerna ändrats från x_i i (1) till $x_i - \bar{x}$.) (3 p)

b) Bestäm medelfelet $d = \sqrt{\widehat{\mathrm{Var}}(\hat{\alpha})}$ för skattningen av metallegeringens hållfasthet. (Ledning: Börja med att ge ett uttryck för $\mathrm{Var}(\hat{\alpha})$. För att skatta denna varians har du hjälp av att först räkna ut kvadratsumman $\mathrm{Kvs(Residual)} = \mathrm{Kvs(Total)} - \mathrm{Kvs(Regression)}$.) (4 p)

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

Uppgift 2

En medicinsk forskargrupp ville undersöka om det fanns några genetiska riskfaktorer för åldersdiabetes. Blodsockerhalten hos 25 diabetespatienter mättes (enhet mmol/l), liksom värdena på tre livsstilsfaktorer och två genetiska riskfaktorer. Man genomförde sedan en multipel linjär regressionsanalys med blodsockerhalt som responsvariabel. I grundmodellen ingick intercept samt både livsstilsfaktorer och genetiska riskfaktorer som förklarande variabler. I hypotesmodellen ingick intercept, och sedan endast livsstilsfaktorerna som förklarande variabler. Feltermerna i regressionsmodellen antogs vara oberoende och normalfördelade med väntevärde 0 och varians σ^2. Resultatet av analysen framgår av följande variansanalystabell:

Variationskälla	Kvs
Regression för hypotesmodell	72.1
Avvikelse mellan grund- och hypotesmodell	48.6
Residual	75.6

a) Testa på nivån 5% om de genetiska riskfaktorerna har en signifikant inverkan på åldersdiabetes, utöver livsstilsfaktorerna. (4 p)

b) Beräkna förklaringsgraden R_1^2 för hypotesmodellen och R_0^2 för grundmodellen. (3 p)

c) Beräkna den justerade (adjusted) förklaringsgraden $R^2_{0,\mathrm{adj}}$ för grundmodellen, genom att först räkna ut två olika skattningar av feltermsvariansen σ^2. (3 p)

Uppgift 3

En plast är en polymerkedja som består av en eller flera olika typer av monomerer. Vid en kemisk processindustri ville man undersöka hur hållfastheten av en viss plasttyp ändrades då proportionen av de tre ingående monomererna varierades kring de standardvärden som användes vid tillverkning av plasten. Man ansatte en multipel linjär regressions modell

$$Y_i = \tilde{\alpha} + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{3i} - \bar{x}_3) + \varepsilon_i,$$

för hållfastheten hos prov nummer $i = 1, \ldots, 24$, där ε_i är oberoende och $N(0, \sigma^2)$-fördelade feltermer, x_{ji} är koncentrationen av monomer j i försök i, och $\bar{x}_j = \sum_{i=1}^{24} x_{ji}/24$. Man varierade koncentrationen av varje monomer på två nivåer (över och under respektive standardvärde), på ett sådant sätt att tre prover togs för alla $2^3 = 8$ nivåkombinationer av de tre monomerernas koncentrationer. Det innebär att designmatrisen för försöket hade ortogonala kolumner. Resultatet av studien sammanfattas i följande variansanalystabell:

Variationskälla	Kvs
Monomer 1	7.1
Monomer 2	6.2
Monomer 3	4.3
Residual	28.2
Total	45.8

a) Utför ett hypotestest på nivån 5%, där grundmodellen har alla tre monomererna som förklarande variabler, medan hypotesmodellen endast har monomer 1 och monomer 2 som förklarande variabler. (Ledning: På grund av ortogonaliteten mellan designmatrisens kolumner, är kvadratsumman för regressionsdelen hos en viss modell, lika med summan av kvadratsummorna för de monomerer som ingår i modellen.) (4 p)

b) Genomför första steget i backward elimination (BE). Det vill säga utgå från den fulla modellen (grundmodellen i a) och visa att en viss förklarande variabel ska tas bort. (Ledning: Deluppgift a) utgör en del av det första steget av BE.) (3 p)

c) Fortsätt till BE-schemats andra steg, efter det första steget i b). Är det så att BE-schemat stannar efter det andra steget eller ska mer än en förklarande variabel ska tas bort? Motivera ditt svar. (Ledning: Grundmodellen i c) är inte densamma som i a) och b), och därför är denna grundmodells Kvs(Residual) ej densamma som i tabellen ovan.) (3 p)

Uppgift 4

Ett läkemedelsföretag misstänkte att två olika mediciner (A och B) hade en gemensam oönskad biverkning, att öka hjärtfrekvensen för de patienter som tog både A och B. För att undersöka saken närmare utgick man från ett stort patientregister, där det både fanns noterat vilka mediciner varje individ tog, och deras dos (ingen dos, låg dos eller hög dos). Man valde ut 18 personer från registret, fördelat på nio grupper som svarade mot alla doskombinationer av A och B. Inom varje grupp valdes två personer ut slumpmässigt bland alla patienter i registret med denna doskombination. Sedan genomfördes en tvåsidig variansanalys typ I, där resultatet sammanfattas i följande variansanalystabell:

Variationskälla	Kvs
Medicin A	9.1
Medicin B	8.4
Samspel	7.7
Residual	9.3

a) Definiera modellen. Förklara de parametrar du inför, och eventuella restriktioner på dem. (3 p)

b) Använd Kvs(Samspel) och Kvs(Residual) för att testa på nivån 5% om samspelet mellan de två medicinera är signifikant. (3 p)

c) Testa på nivån 5% (med ett test) om de kombinerade huvudeffekterna av Medicin A eller B är signifikant. Låt din analys bero av svaret i b), så att ett eventuellt icke-signifikant samspel tas bort från den förklarande delen av modellen. (Ledning: För att studera den gemensamma effekten av ortogonala variationskällor kan motsvarande kvadratsummor slås ihop.) (4 p)

Uppgift 5

En stationär AR(1)-process definieras enligt

$$X_t = \phi X_{t-1} + \varepsilon_t$$

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

a) Beräkna kovariansfunktionen $\gamma_k = \mathrm{Cov}(X_t, X_{t+k})$ och korrelationsfunktionen $\rho_k = \mathrm{Corr}(X_t, X_{t+k})$ för $k = 0, 1, 2, \ldots$ (Ledning: Börja med att beräkna γ_0 och härled sedan ett rekursivt uttryck för γ_k för $k = 1, 2, \ldots$. Därefter kan ρ_k bestämmas. Du kan utan bevis utnyttja att $\mathrm{Cov}(X_t, \varepsilon_{t+k}) = 0$ för $k = 1, 2, \ldots$.) (4 p)

b) Anta att T mätningar $\boldsymbol{X}_T = (X_1, \ldots, X_T)$ registrerats av processen. Härled ett uttryck för $\hat{X}_{T+k} = E(X_{T+k}|\boldsymbol{X}_T)$ för $k = 1, 2, 3, \ldots$. (Ledning: Vid beräkning av $\hat{X}_{T+k}$ kan du utnyttja att en AR(1)-process är en Markovprocess och alltså ersätta $\boldsymbol{X}_T$ med X_T i defnitionen av $\hat{X}_{T+k}$.) (3 p)

c) Bestäm prediktionsfelsvariansen $\sigma_k^2 = \mathrm{Var}(X_{T+k} - \hat{X}_{T+k})$ för $k = 1, 2, 3, \ldots$, där $\hat{X}_{T+k}$ är prediktionen av X_{T+k} från b). (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