

Tentamen för kursen

Introduktion till övervakad statistisk inlärnin

25 november 2025 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 tentamens-tillfället. Tabell över F-kvantiler återfinns nedan. Det gäller även att $\chi^2_{0.05}(1) \approx 3.8$ och $\chi^2_{0.05}(2) \approx 5.99$.

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

Vattnet i en viss region innehåller två skadliga bakterietyper A och B. Halten av B-bakterier (enhet: mg/l) varierar mellan områden i regionen, medan halten A-bakterier kan anses vara konstant. Biologen Lisa har en utrustning som medger mätning av den totala bakteriehalten för vattenprover som genomgått ett visst reningsfilter. Detta filter eliminerar andelen $1 - \beta$ av alla B-bakterier, medan alla typ A-bakterier passerar igenom filtret. Lisa analyserade vattenprover vid 16 olika fältstationer inom regionen och ställde upp en enkel linjär regressionsmodell

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, 16, \quad (1)$$

för sambandet mellan bakteriehalten Y_i för provet vid station $i = 1, \dots, 16$ efter rening och andelen x_i av typ B-bakterier vid station i före rening ($x_1, \dots, x_{16}$ hade tidigare bestämts med hjälp av en noggrannare mätutrustning). Således svarar α mot halten av typ A-bakterier i hela regionen. Lisa

antog vidare att ε_i är oberoende och $N(0, \sigma^2)$ -fördelade feltermer. Resultatet från de 16 mätningarna sammanfattades i form av följande fem summer;

$$\begin{aligned}\sum_{i=1}^{16} x_i &= 168.0, \\ \sum_{i=1}^{16} (x_i - \bar{x})^2 &= 45.0, \\ \sum_{i=1}^{16} Y_i &= 312.0, \\ \sum_{i=1}^{16} Y_i(x_i - \bar{x}) &= 9.0, \\ \sum_{i=1}^{16} (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 &= 19.0,\end{aligned}$$

där $\hat{\alpha}$ och $\hat{\beta}$ är minsta kvadratskattningarna av halten A-bakterier respektive andelen B-bakterier som inte elimineras av filtret.

a) Beräkna $\hat{\alpha}$ och $\hat{\beta}$. (Ledning: Analysera först en centrerad regressionsmodell med intercept $\tilde{\alpha}$, där de förklarande variablerna ändrats från x_i i den icke-centrerade parametreringen (??), till $x_i - \bar{x}$. Utnyttja sedan att $\hat{\alpha}$ är en linjärkombination av $\hat{\tilde{\alpha}}$ och $\hat{\beta}$.) (3 p)

b) Bestäm medelfelet $d = \sqrt{\widehat{\text{Var}}(\hat{\alpha})}$ för skattningen $\hat{\alpha}$ av det icke-centrerade interceptet α . (Ledning: Börja med att bestämma $\text{Var}(\hat{\alpha})$, där du utnyttjar sista meningen från ledningen i 1a.) (4 p)

c) Ange ett 95% konfidensintervall för α . (Ledning: Du kan använda formeln $t_{0.025}(f) = \sqrt{F_{0.05}(1, f)}$ för kvantilen hos en t -fördelning med f frihetsgrader, där F -fördelningens kvantil fås ur tabell.) (3 p)

Uppgift 2

Vid en stålindustri framställs en legering som består av tre metaller. Man vill ta reda på hur känslig legeringens hårdhet är för små variationer kring nuvarande värden av de tre ingående metallernas koncentration. Totalt genomförs $N = 20$ experiment $i = 1, \dots, 20$ där hårdheten Y_i och koncentrationerna x_{1i} , x_{2i} , x_{3i} av de tre metallerna registreras. Man antar att metallerna påverkar legeringens hårdhet oberoende av varandra. Därför bortses från samspel mellan metallernas inverkan på legeringens hårdhet och olika linjära regressionsmodeller med ingen, en, två eller tre metallkoncentrationer som förklarande variabler jämförs. Dessutom är försöket upplagt så att effekterna av de olika förklarande variablerna är ortogonala, det vill säga

$$\sum_{i=1}^{20} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) = 0 \quad (2)$$

för alla $1 \leq j < k \leq 3$, med $\bar{x}_j = \sum_i x_{ji}/20$.

Variationskälla	Kvs
Metall 1	3.1
Metall 2	5.1
Metall 3	8.9
Residual	25.0
Totalt	42.1

a) För den *fullständiga modellen* med alla tre metaller får man ovanstående variansanalystabell. Fyll i antalet frihetsgrader för alla fyra variationkällorna och för Totalt. (2 p)

b) Genomför första steget i framåtinkludering (Forward Selection, FS). Undersök alltså om någon förklarande variabel ska tas med. Signifikansnivån väljs till 5%. (Ledning: På grund av ortogonaliteten (??) mellan de förklarande variablerna fås kvadratsumman för avvikelserna mellan en grund- och en hypotesmodell som summan av kvadratsummorna (i tabellen ovan) för de metaller som ingår i grundmodellen men inte i hypotesmodellen. För varje delmodell av den fullständiga modellen med alla tre metaller, så fås kvadratsumman för variationskällan Residual för delmodellen som summan av kvadratsumman Residual för den fullständiga modellen och kvadratsummorna för de metaller som *inte* ingår i delmodellen.) (4 p)

c) Stannar FS-schemat efter 2b)? Motivera ditt svar. (4 p)

Uppgift 3

Vid ett medicinskt laboratorium mäter man diametern hos röda blodkroppar med hjälp av ett mikroskop (enhet: μm). Man vill uppskatta hur mycket blodkropparna hos en individ varierar i storlek. Vid ett tillfälle valde man ut 6 blodkroppar från en patient och genomförde 4 mätningar på varje blodkropp. De uppmätta diametrarna $Y_{i1}, \dots, Y_{i4}$ för blodkropp $i = 1, \dots, 6$ antas följa en ensidig variansanalysmodell

$$Y_{ij} = \mu + \delta_i + \varepsilon_{ij}$$

av typ II, med väntevärde μ , oberoende effekter $\delta_i \sim N(0, \sigma_\delta^2)$ av blodkropparna, samt oberoende mätfel $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. Resultatet sammanfattas i följande tabell, i form av stickprovsmedelvärden $\bar{Y}_i$ och stickprovsvarianser s_i^2 för mätningarna på respektive blodkropp:

Blodkropp i	$\bar{Y}_i$	s_i^2
1	7.5	0.020
2	7.1	0.015
3	8.2	0.025
4	7.4	0.018
5	7.9	0.022
6	7.5	0.014

- a)** Skatta de två varianskomponenterna σ_δ^2 och σ_ε^2 . (Ledning: Börja med att skatta σ_ε^2 med hjälp av stickprovsvarianserna. Du kan även utnyttja att $\sum_{i=1}^6 (\bar{Y}_i - \bar{Y}_\cdot)^2 / 5 = 0.152$.) (5 p)
- b)** Ange ett 95% konfidensintervall för μ . (Ledning: Du har användning av $\bar{Y}_\cdot = (\bar{Y}_1 + \dots + \bar{Y}_6) / 6 = 7.6$. Vid beräkningen av medelfelet för skattningen av μ kan du utnyttja räkningarna i 3a.) (5 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

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t, \quad (3)$$

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

- a)** Låt $\rho_k = \text{Corr}(Y_t, Y_{t+k}) = \text{Corr}(X_t, X_{t+k})$, $k = 0, 1, 2, \dots$ 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 (??) och att $\text{Cov}(X_t, \varepsilon_{t+k}) = 0$ för $k > 0$, för att ge uttryck för $\gamma_1 = \text{Cov}(X_t, X_{t+1})$ och $\gamma_2 = \text{Cov}(X_t, X_{t+2})$ som innefattar $\gamma_0 = \text{Var}(X_t)$. Uttryck därefter ρ_k med hjälp av γ_k och γ_0 .) (5 p)

- b)** Anta att vi har data $y_1, \dots, y_T$ från tidsserien i 4a). Definiera skattningen $\hat{\rho}_k$ av autokorrelationsfunktionen för $0 \leq k < T$. (3 p)

- c)** Anta att vi vill testa nollhypotesen

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

Vi ser från 4a) att nollhypotesen medför $\rho_1 = \rho_2 = 0$. För att testa om H_0 är sann kan man bilda teststorheten

$$Q = T(\hat{\rho}_1^2 + \hat{\rho}_2^2),$$

där $\hat{\rho}_1$ och $\hat{\rho}_2$ är skattningarna av ρ_1 och ρ_2 från 4b), baserat på T datapunkter. Konstruera ett test som för stora värden på T approximativt har signifikansnivån 5%. (Ledning: Du får utan bevis använda att då T är stor så gäller approximativt under H_0 att $\hat{\rho}_1$ och $\hat{\rho}_2$ är oberoende, samt att $\sqrt{T}\hat{\rho}_k \sim N(0, 1)$.) (2 p)

Uppgift 5

Anta att vi har sex observationer

$$Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, \dots, 6, \quad (4)$$

av en responsvariabel Y_i och två förklarande variabler/kovariater x_{1i} och x_{2i} , för en linjär modell utan intercept, där β_1 och β_2 är effektparametrarna för de två kovariaterna och där feltermerna $\varepsilon_i \sim N(0, \sigma^2)$ antas oberoende. Värdena på de förklarande variablerna för de sex försökspunkterna återfinns i följande tabell:

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

Med hjälp av detta dataset vill man prediktera värdet på en ny observation

$$Y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon = \mu + \varepsilon,$$

där designpunkten (x_1, x_2) är känd, och där $\varepsilon \sim N(0, \sigma^2)$ är oberoende av feltermerna ε_i i (??).

a) Härled den tvådimensionella normalfördelningen för minsta kvadrat-skattningen $(\hat{\beta}_1, \hat{\beta}_2)^T$ av $(\beta_1, \beta_2)^T$ baserad på modellen i (??). (3 p)

b) Responsvariabeln för den nya observationen kan predikteras med $\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Använd 5a) för att bestämma medelkvadratfelet

$$\text{MSEP} = E[(Y - \hat{Y})^2] = E[(Y - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2)^2]$$

för prediktionen av den nya observationen. (Ledning: Ditt svar kommer att bero på x_1 , x_2 och σ^2 . Du kan utnyttja att prediktionsfelet $Y - \hat{Y}$ kan uttryckas med hjälp av ε och skattningsfelen $\hat{\beta}_1 - \beta_1$ och $\hat{\beta}_2 - \beta_2$ för de två effektparametrarna.) (3 p)

c) Anta att vi istället anpassar datamaterialet med sex observationer till en modell där endast den första kovariaten finns med, dvs

$$Y_i = \beta_1 x_{1i} + \epsilon_i, \quad i = 1, \dots, 6, \quad (5)$$

där $\epsilon_i = \varepsilon_i + \beta_2 x_{2i}$ ses som feltermer. Visa att man i denna modell får samma minsta kvadrat-skattning $\hat{\beta}_1$ av β_1 som i 5a), och alltså en prediktor $\hat{\beta}_1 x_1$ av den nya observationen Y . Använd sedan detta till att bestämma medelkvadratfelet

$$\text{MSEP}_1 = E[(Y - \hat{\beta}_1 x_1)^2]$$

för prediktion av Y med hjälp av modellen (??), som endast baseras på kovariat 1. För vilka värden på (x_1, x_2) och $(\beta_1, \beta_2, \sigma^2)$ får man lägre prediktionsfel genom att använda den mindre (och felaktiga) modellen (??) jämfört med den korrekta modellen (??)? (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