STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

18 augusti 2021 8–15

Examinator: Ola Hössjer. Kan nås under skrivtiden via mobil (070/672 12 18) eller mejl (ola@math.su.se).

Inlämning: Lösningar mejlas till examinator senast kl 15 (eller 16 för studenter med förlängd skrivtid) i form av en pdf-fil. Denna fil kan antingen innehålla inscannade och handskrivna lösningar eller lösningar som skrivits ned i en ordbehandlare (t ex LaTeX).

Återlämning: Meddelas via mejl.

Tillåtna hjälpmedel: Miniräknare och formelsamling, samt lärobok och andra skriftliga informationskällor. Tabell över F-kvantiler återfinns nedan. Det gäller även att $\chi^2_{0.05}(1) \approx 3.8$. Det är inte tillåtet att ta hjälp av andra personer.

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 0

Skriv en försäkran att du löst alla uppgifter självständigt. Detta krävs för att tentan ska rättas.

(0 p)

Uppgift 1

En grupp epidemiologer studerade sambandet mellan kaffekonsumtion och hjärtfrekvens vid vila. De noterade hjärtfrekvens Y_i (enhet: slag/minut)

och kaffekonsumtion x_i (enhet: koppar per dag) för $i = 1, \ldots, 20$ personer. Vidare satte de upp en enkel linjär regressionsmodell

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, 20, \tag{1}$$

för sambandet mellan kaffedrickande och hjärtfrekvensen vid vila. Här är ε_i oberoende och $N(0, \sigma^2)$-fördelade feltermer, α svarar mot den genomsnittliga hjärtfrekvensen för personer som inte dricker kaffe, medan β anger hur mycket hjärtfrekvensen ändras då kaffekonsumtionen ökas med en kopp per dag. Resultatet från studien sammanfattas med summorna

$$\begin{aligned} \textstyle\sum_{i=1}^{20} x_i &= 60.0, \\ \textstyle\sum_{i=1}^{20} (x_i - \bar{x})^2 &= 85.0, \\ \textstyle\sum_{i=1}^{20} Y_i &= 1580.0, \\ \textstyle\sum_{i=1}^{20} Y_i (x_i - \bar{x}) &= 160.2, \\ \textstyle\sum_{i=1}^{20} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 &= 301.0, \end{aligned}$$

där $\bar{x} = \sum_i x_i / 20$.

a) Beräkna minsta kvadrat-skattningen $\hat{\alpha}$ av α. (Ledning: Skatta först parametrarna i en centrerad regressionsmodell $Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i$, 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 i a). (4 p)

c) Ange ett 95% konfidensintervall för α. (Ledning: Erforderlig t-kvantil fås som kvadratroten av en F-kvantil som återfinns i tabellen längst ned i denna skrivning.) (3 p)

Uppgift 2

I en betongfabrik undersöks hur betongens hållfasthet ändras då mängden vatten och cement varierar kring de värden som för närvarande används vid betongproduktionen. Koncentrationerna av övriga ingredienser (ballast och kemiska tillsatsmedel) hålls däremot konstanta. Totalt genomförs $N = 25$ experiment $i = 1, \ldots, 25$ där vatten x_{1i} och cement x_{2i} registreras, liksom hållfastheten Y_i (genom att den stelnade betongen får genomgå olika accelererade stresstester). I det område inom vilket de förklarande variablerna varierar så antar man att vatten- och cementmängden påverkar hållfastheten linjärt och additivt, utan något samspel. Därför jämförs linjära regressionsmodeller med ingen, en eller två förklarande variabler. Försöket är upplagt så att de två förklarande variablerna är ortogonala, det vill säga

$$\sum_{i=1}^{25} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = 0,$$

med $\bar{x}_j = \sum_{i=1}^{25} x_{ji} / 25$. För den *fullständiga modellen* med två kovariater får man följande variansanalystabell:

Variationskälla	Kvs
Vatten	4.5
Cement	4.4
Residual	22.0
Totalt	30.9

a) 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 avvikelsen mellan en grund- och en hypotesmodell som kvadratsumman (i tabellen ovan) för den förklarande variabel som ingår i grundmodellen men inte i hypotesmodellen.) (5 p)

b) Genomför första steget i bakåteliminering (Backward elimination (BE). Diskutera eventuella skillnader gentemot resultatet i a). (5 p)

Uppgift 3

Fältbiologen Eva undersöker hur brunröta i träd (reduktion av mängden lignin) påverkas av mängden av två svampsorter. Som försöksmaterial använder hon sig av 30 träd i ett växthus. Dessa träd tillhör samma art och är lika gamla. Eva injicerar olika koncentrationer av de två svampsorterna på respektive träd och mäter sedan hur mycket ligninmängden Y på trädet har minskat ett år senare. Detta sker enligt en tvåsidig variansanalys typ 1, där koncentrationen av svampsort 1 (faktor 1) varieras på tre nivåer medan koncentrationen av svampsort 2 (faktor 2) varieras på fem nivåer. För varje nivåkombination av de två faktorerna analyseras två träd.

a) Formulera den tvåsidiga variansanalysmodellen matematiskt, där både huvudeffekter och samspel ingår. (3 p)

b) En variansanalystabell från försöket har följande utseende:

Variationskälla	Kvs
Svampsort 1	8.5
Svampsort 2	15.8
Samspel	15.2
Inom celler	14.0
Total	53.5

Testa på nivån 5% om det finns något signifikant samspel mellan hur koncentrationen av de två svampsorterna tillsammans påverkar ligninreduktionen. (3 p)

c) Testa på nivån 5% om svampsort 1 har en signifikant påverkan på ligninreduktionen. Variationskällan samspel tas med för att skatta feltermernas varians eller ej, beroende på om samspelet i deluppgift b) inte är eller är signifikant. Kommentera även vilken inverkan valet av skattning av σ^2 har. (4 p)

Uppgift 4

Lasse är en statistikintresserad bagare som vill undersöka hur koncentrationen av surdeg (S), vatten (V) och mjöl (M) påverkar volymen Y hos den jästa degen efter 10 timmar. Han genomför ett 2^3-försök utan replikat, där volymen av den jästa degen bestäms då mängden surdeg, vatten och mjöl varieras på en låg (-) och en hög (+) nivå. Låt Y_{ijk} beteckna volymen hos den jästa degen (enhet: dm^3) då S, V och M är på nivåerna $i, j, k \in \{-, +\}$. Tabellerna nedan visar var sitt fraktionellt 2^{3-1}-försök, som båda utgör delar av det fullständiga 2^3-försöket.

S	V	M	Y_{ijk}
-	-	-	10.5
+	-	+	15.5
-	+	-	12.5
+	+	+	18.5

S	V	M	Y_{ijk}
-	-	+	11.5
-	+	-	12.5
+	-	-	14.5
+	+	+	18.5

a) Bestäm kopplingsschemat för respektive försök. (Ledning: Börja med att fylla ut respektive teckentabell med kolumner för SV, SM, VM och SVM.) (3 p)

b) Anta att alla interaktioner av ordning 2 och 3 mellan de tre faktorerna kan försummas, och ansätt en additiv modell

$$Y_{ijk} = \mu + \bar{S} \cdot i + \bar{V} \cdot j + \bar{M} \cdot k + \varepsilon_{ijk}, \tag{2}$$

där μ anger intercept, och $\bar{S}, \bar{V}, \bar{M}$ effekten av respektive faktor. Feltermerna ε_{ijk} är oberoende och normalfördelade med väntevärde 0 och varians σ^2. För vilket av de två fraktionella försöken ovan kan minsta kvadrat-skattningar av de tre huvudeffekterna $\bar{S}, \bar{V}, \bar{M}$ beräknas? Beräkna dessa skattningar $\hat{S}, \hat{V}, \hat{M}$ för det försök du valde. Motivera även om feltermsvariansen σ^2 skattas för modellen (2). (4 p)

c) Låt

$$Y_{ijk} = \mu + \bar{S} \cdot i + \bar{V} \cdot j + \epsilon_{ijk} \tag{3}$$

vara den additiva modell där huvudeffekten för mjöl tagits bort i (2) och inkluderats i den nya feltermerna $\epsilon_{ijk} = \bar{M} \cdot k + \varepsilon_{ijk}$, som antas oberoende och $N(0, \tilde{\sigma}^2)$-fördelade. Använd skattningarna i b) för att bestämma förklaringsgraden R^2 hos modellen (3), för det dataset du valde i b). (Ledning: För varje faktor $F \in \{S, V, M\}$ i b) gäller $\mathrm{Kvs}(F) = 4\hat{F}^2$. Dessutom är dessa tre faktorer ortogonala, så att kvadratsumman för en variationskälla med flera faktorer är summan av de ingående faktorernas kvadratsummor.) (3 p)

Uppgift 5

Låt

$$Y_{ij} = \mu + \delta_i + \varepsilon_{ij} = \mu + \epsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n, \tag{4}$$

beteckna observationer från en ensidig variansanalys typ II, där $\mu = E(Y_{ij})$ är väntevärdet för responsvariabeln, $\delta_i \sim N(0, \sigma_\alpha^2)$ anger inverkan från faktorns olika nivåer medan $\varepsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ är feltermer. Vidare antas att alla δ_i och ε_{ij} vara oberoende av varandra. Antalet observationer är $N = kn$, modellens enda regressionsparameter är μ och fluktuationerna hos Y_{ij} kring μ beskrivs av $\epsilon_{ij} = \delta_i + \varepsilon_{ij}$.

a) Visa att (4) kan skrivas som en allmän linjär modell

$$\boldsymbol{Y} = \boldsymbol{A}\mu + \boldsymbol{\epsilon},$$

där $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1n}, Y_{21}, \ldots, Y_{kn})^T$ är en responsvektor av dimension $N \times 1$, $\boldsymbol{A}$ en designmatris av ordning $N \times 1$ och $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{1n}, \epsilon_{21}, \ldots, \epsilon_{kn})^T$ en $N \times 1$-vektor med väntevärde $E(\boldsymbol{\epsilon}) = \mathbf{0}$ och kovariansmatris $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$, som har en blockdiagonal struktur

$$\boldsymbol{\Sigma} = \mathrm{BDiag}(\boldsymbol{\Lambda}, \ldots, \boldsymbol{\Lambda}) = \begin{pmatrix} \boldsymbol{\Lambda} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda} & \ldots & \mathbf{0} \\ \vdots & & \ldots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Lambda} \end{pmatrix},$$

med identiska $k \times k$-matriser $\boldsymbol{\Lambda}$ längs diagonalen. Ange speciellt $\boldsymbol{A}$ och $\boldsymbol{\Lambda}$. (Ledning: Du kommer ha nytta av att beräkna kovariansen mellan olika ϵ_{ij}.) (3 p)

b) Man kan visa att den generaliserade minsta kvadrat-skattningen av μ ges av $\hat{\mu} = (\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$, där $\boldsymbol{\Sigma}^{-1} = \mathrm{BDiag}(\boldsymbol{\Lambda}^{-1}, \ldots, \boldsymbol{\Lambda}^{-1})$ är inversen av $\boldsymbol{\Sigma}$. Visa med hjälp av detta att

$$\mathrm{Var}(\hat{\mu}) = (\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}$$

är variansen för parameterskattningen. (Ledning: Börja med att skriva skattningen av μ på formen $\hat{\mu} = \boldsymbol{B}\boldsymbol{Y}$ för lämpligt vald matris $\boldsymbol{B}$.) (2 p)

c) Använd a) och b) för att förenkla uttrycken för $\hat{\mu}$ och $\mathrm{Var}(\hat{\mu})$ så långt som möjligt, genom att först förenkla uttrycken för $\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}$ och $\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}$. (Ledning: Om $\mathbf{1}_n = (1, \ldots, 1)^T$ är en kolumnvektor med n ettor och $\boldsymbol{I}_n$ enhetsmatrisen av ordning n så gäller att $(a\mathbf{1}_n\mathbf{1}_n^T + b\boldsymbol{I}_n)^{-1} = -a\mathbf{1}_n\mathbf{1}_n^T/(nab + b^2) + \boldsymbol{I}_n/b$ för alla a och b sådana att $nab + b^2 \neq 0$.) (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