STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

25 oktober 2019 9–14

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

Några statistiker ville studera sambandet mellan nattsömn och korttidsminne. Totalt deltog 30 personer i undersökningen, där minnesförmågan Y_i hos person i poängsattes utifrån ett antal tester. Här svarar ett värde på Y_i under 20, mellan 20 och 25, samt över 25 mot ett dåligt, medelgott respektive bra korttidsminne. Varje deltagare fick även ange hur många timmar x_i han eller hon sovit natten innan (avrundat till heltal). Forskarna ställde upp en enkel linjär regressionsmodell

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 30, \tag{1}$$

för minnesförmågan hos deltagarna, där $\bar{x} = \sum_{i=1}^{30} x_i/30$ är deras genomsnittliga antal timmars nattsömn. Vidare antog statistikerna att feltermerna ε_i är oberoende och normalfördelade med väntevärde 0 och varians σ^2. Man sammanfattade undersökningen genom att dela upp deltagarna i 5 grupper beroende på hur länge de sovit (x_i lika för alla personer i samma grupp). Resultatet framgår i följande tabell

Timmars nattsömn	Antal	Medel
5	4	19.0
6	6	22.0
7	10	24.0
8	6	25.5
9	4	26.5
Totalt	30	

där Medel för en viss rad anger medelvärdet av alla Y_i för personer med ett visst antal timmars nattsömn.

a) Beräkna minsta kvadrat-skattningarna $\hat{\alpha}$ och $\hat{\beta}$ av α och β. (Ledning: Du kan utnyttja att $\sum_{i=1}^{30}(x_i - \bar{x})^2 = 44$ och $\sum_{i=1}^{30}(x_i - \bar{x})Y_i = 81$.) (3 p)

b) Bestäm den tvådimensionella fördelningen för $(\hat{\alpha}, \hat{\beta})$, uttryckt med hjälp av α, β och σ^2. (2 p)

c) En variansanalystabell från försöket innehöll kvadratsumman för variationskällan Residual (Kvs(Residual) = 550). Beräkna med hjälp av denna information en väntevärdesriktig skattning av σ^2. (2 p)

d) Använd a-c för att bestämma ett 95% konfidensintervall för den förväntade minnesförmågan $\mu = E(Y)$ hos en person som sov 6.5 timmar natten innan undersökningen gjordes. (3 p)

Uppgift 2

En grupp epidemiologier ville utröna hur rökning och graden av fysisk aktivitet tillsammans påverkade syreupptagningsförmågan. Man undersökte totalt 24 personer. Varje person fick ange hur ofta han eller hon rökte, uppdelat på tre nivåer (aldrig/ibland/varje dag), medan den fysiska aktiviteten hade två nivåer (låg och hög). Studien var balanserad såtillvida att 4 personer ingick i patientgruppen för varje nivåkombination av rökning och fysisk aktivitet.

a) Formulera en tvåsidig variansanalysmodell där båda faktorerna rökning och fysisk aktivitet är systematiska, och där samspelet mellan dessa båda faktorer ingår. (3 p)

b) En variansanalystabell från försöket har följande utseende:

Variationskälla	Kvs
Rökning	10.0
Fysisk aktivitet	6.0
Samspel	5.5
Inom celler	19.5
Total	41.0

Testa på nivån 5% om det finns något signifikant samspel mellan hur rökning och fysisk aktivitet tillsammans påverkar syreupptagningsförmågan. (3 p)

c) Testa på nivån 5% om rökning har en signifikant påverkan på syreupptagningsförmågan. Variationskällan samspel tas med för att skatta feltermernas varians eller ej, beroende på om samspelet i deluppgift b) inte är eller är signifikant. (4 p)

Uppgift 3

En forskargrupp undersökte utbytet vid en viss kemisk reaktion. Man genomförde ett 2^3-försök utan replikat, där reaktionsutbytet studerades då katalysatorkoncentration C, tryck P och temperatur T varierades på en låg (-) och en hög (+) nivå. Låt Y_{ijk} beteckna reaktionsutbytet vid försöket då C, P och T valdes 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.

C	P	T	Y_{ijk}
+	-	-	3.5
-	+	-	4.5
-	-	+	6.5
+	+	+	10.5

C	P	T	Y_{ijk}
-	-	-	2.5
+	-	+	7.5
-	+	-	4.5
+	+	+	10.5

a) Bestäm kopplingsschemat för respektive försök. (3 p)

b) Vi antar nu att alla interaktioner av ordning 2 och 3 mellan de tre faktorerna kan försummas, och ansätter en additiv modell

$$Y_{ijk} = \mu + \bar{C} \cdot i + \bar{P} \cdot j + \bar{T} \cdot k + \varepsilon_{ijk},$$

där μ anger försökens totala väntevärde, och $\bar{C}, \bar{P}, \bar{T}$ effekten av respektive faktor. Feltermerna $\varepsilon_{ijk} \sim N(0, \sigma^2)$ antas vara oberoende. För vilket av de två fraktionella försöken ovan kan minsta kvadrat-skattningar av de tre huvudeffekterna $\bar{C}, \bar{P}, \bar{T}$ beräknas? Beräkna dessa skattningar $\hat{C}, \hat{P}, \hat{T}$ för det försök du valde. (3 p)

c) Låt

$$\mu_{ijk} = \mu + \bar{C} \cdot i + \bar{P} \cdot j + \bar{T} \cdot k$$

vara det förväntade reaktionsutbytet då de tre faktorerna är på nivå i, j, k. Speciellt anger $\Delta = \mu_{+++} - \mu_{---}$ hur mycket reaktionsutbytet ändras då alla tre faktorerna ändras från den låga till den höga nivån. Beräkna motsvarande skattning $\hat{\Delta}$, och dess varians $\mathrm{Var}(\hat{\Delta})$, för det fraktionella försök du valde i deluppgift b). Går det att skatta denna varians utifrån detta fraktionella försök? (Ledning: Δ kan skrivas som en linjärkombination av modellens regressionsparametrar $\boldsymbol{\theta} = (\mu, \bar{C}, \bar{P}, \bar{T})^T$. Bestäm kovariansmatrisen för skattningen av $\boldsymbol{\theta}$.) (4 p)

Uppgift 4

Ett företag genomför en enkel bestämning av personers genetiska härkomst från två regioner 1 och 2. Syftet är att för varje person som lämnat in ett blodprov uppskatta proportionerna β_1 och β_2 av hans eller hennes DNA som härrör från respektive region, samt den resterande proportionen $1-\beta_1-\beta_2$ av DNA som svarar mot ett ursprung från andra regioner (=region 0). Metoden går ut på att man hittat $N = 4$ grupper av genvarianter som förekommer i följande kända proportioner p_{ji} i region j för grupp i:

	Region j		
Grupp i	0	1	2
1	0.5	0	0
2	0.5	1	0
3	0.5	0	1
4	0.5	1	1

För en viss person bestäms proportionen

$$Z_i = 0.5(1 - \beta_1 - \beta_2) + \beta_1 p_{1i} + \beta_2 p_{2i} + \varepsilon_i, \tag{2}$$

av genvarianterna i grupp $i = 1, 2, 3, 4$ som förekommer i hans eller hennes DNA-prov, där $\varepsilon_i \sim N(0, \sigma^2)$ antas vara oberoende feltermer. Här kan alltså p_{1i} och p_{2i} avläsas ur de två högra kolumnerna från tabellen ovan. Genom att införa $x_{ji} = p_{ji} - 0.5$ och $Y_i = Z_i - 0.5$ kan (2) skrivas som en multipel linjär regressionsmodell

$$Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, 2, 3, 4, \tag{3}$$

med två förklarande variabler och utan intercept. Även om β_1 och β_2 tolkas om proportionen av härkomsten från region 1 och 2, görs inga restriktioner i (3) att dessa två parametrar ska ligga mellan 0 och 1.

a) Kalle skickar in sitt DNA-prov till företaget och får följande proportioner av genvarianterna uppmätta för de fyra grupperna:

Grupp i	Z_i
1	0.23
2	0.41
3	0.62
4	0.74

Bestäm minsta-kvadrat-skattningen $(\hat{\beta}_1, \hat{\beta}_2)^T$ av Kalles härkomst. (Ledning: Börja med att räkna ut Y_i, x_{1i} och x_{2i}.) (2 p)

b) Bestäm kovariansmatrisen för skattningen i a) och därefter variansinflationsfaktorn $\text{VIF}(\hat{\beta}_1)$ för skattningen av graden av härkomst från region 1. (4 p)

c) Bestäm en tvådimensionell konfidensregion för $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ med konfidensgrad 0.95. (Ledning: Börja med att skatta σ^2. Utnyttja att $\sum_{i=1}^4 Y_i^2 = 0.153$ kan delas upp i tre kvadratsummor, varav två ges av Kvs(Region j) = $\hat{\beta}_j^2 \sum_{i=1}^4 x_{ji}^2$ för $j = 1, 2$ och den tredje är residualernas kvadratsumma.) (4 p)

Uppgift 5

En multipel linjär regressionsmodell

$$\begin{aligned} Y_i &= \alpha + \beta_1(x_{1i} - \bar{x}_1) + \ldots + \beta_m(x_{mi} - \bar{x}_m) + \varepsilon_i \\ &= \mu_i + \varepsilon_i \end{aligned} \tag{4}$$

uttrycker sambandet mellan responsvariabeln Y_i och de m förklarande variablerna $x_{1i}, \ldots, x_{mi}$ för ett antal individer $i = 1, \ldots, N$, där $\bar{x}_j = \sum_{i=1}^N x_{ji}/N$ och $\varepsilon_i \sim N(0, \sigma^2)$ är oberoende feltermer. Man vill testa om en viss förklarande variabel j har någon effekt på responsvariabeln genom att testa grundmodellen (4) mot hypotesmodellen $H_0 : \beta_j = 0$.

a) Låt R_0^2 och R_1^2 vara förklaringsgraden för grund- respektive hypotesmodellen. Definiera R_0^2 och R_1^2 med hjälp av $\hat{\mu}_i$, $\hat{\hat{\mu}}_i$ och Y_i för alla observationer i, samt med $\bar{Y} = \sum_i Y_i/N$. Här är

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta}_1(x_{1i} - \bar{x}_1) + \ldots + \hat{\beta}_m(x_{mi} - \bar{x}_m)$$

skattningen av μ_i under grundmodellen baserat på minsta kvadrat-skattningar av interecept och effektparametrar, samt $\hat{\hat{\mu}}_i$ motsvarande skattning av μ_i för hypotesmodellen. (3 p)

b) Skillnaden i förklaringsgrad mellan de två modellerna, $R_0^2 - R_1^2$, är ett mått på hur mycket bättre grundmodellen anpassar sig till det givna datasetet. Visa att $R_0^2 - R_1^2$ kan uttryckas med hjälp av $\hat{\mu}_i - \hat{\hat{\mu}}_i$ och Y_i för alla observationer, samt med $\bar{Y}$. (Ledning: Utyttja att vektorn $\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}} = (\hat{\mu}_i - \hat{\hat{\mu}}_i;\, i = 1, \ldots, N)^T$ är ortogonal mot det underrum som spänns upp av hypotesmodellen.) (2 p)

c) Låt $\hat{\hat{x}}_{ji}$ vara en uppskattning av x_{ji} med hjälp av de övriga $m - 1$ förklarande variablerna för observation i. Med andra ord så betraktar man x_{ji} som stokastisk - en responsvariabel i en multipel regressionsmodell med intercept och de övriga $m - 1$ förklarande variablerna som kovariater. I denna modell är $\hat{\hat{x}}_{ji}$ en skattning av $E(x_{ji})$. Visa att

$$\hat{\mu}_i = \hat{\beta}_j(x_{ji} - \hat{\hat{x}}_{ji}) + \hat{\hat{\mu}}_i, \quad i = 1, \ldots, N.$$

Använd sedan detta samband och deluppgift b) för att uttrycka $R_0^2 - R_1^2$ med hjälp av minsta kvadrat-skattningen $\hat{\beta}_j$ av β_j för grundmodellen. (Ledning: Betrakta delrummmen av R^N som spänns upp av hypotes- respektive grundmodellerna. Utnyttja ortogonalitetsegenskaper hos vektorn $\boldsymbol{x}_j - \hat{\hat{\boldsymbol{x}}}_j = (x_{j1} - \hat{\hat{x}}_{j1}, \ldots, x_{jN} - \hat{\hat{x}}_{jN})^T$, samt att $\hat{\boldsymbol{\mu}} = \boldsymbol{A}\hat{\boldsymbol{\theta}}$, där designmatrisen $\boldsymbol{A}$ har en kolumn $\boldsymbol{x}_j - \bar{\boldsymbol{x}}_j = (x_{j1} - \bar{x}_j, \ldots, x_{jN} - \bar{x}_j)^T$, och där $\hat{\boldsymbol{\theta}}$ är minsta kvadratskattningen av regressionsparametrarna för grundmodellen.) (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