STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

19 augusti 2020 9–17

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 17 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 kurshemsidan, webbaserat kursforum eller per 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 viss medicin syftar till att sänka det diastoliska (undre) blodtrycket. I en klinisk studie testades medicinen på 20 patienter med högt blodtryck, där patient $i = 1, \ldots, 20$ fick dosen x_i gram per dag av under en vecka,

varefter blodtryckssänkningen Y_i (enhet: mm Hg) registrerades. En grupp statistiker på läkemedelsföretaget ansatte en linjär regressionsmodell

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 20, \tag{1}$$

för sambandet mellan x_i och Y_i, där $\bar{x} = \sum_{i=1}^{20} x_i/20$ är den genomsnittliga dosen för patienterna i studien, medan ε_i antas vara oberoende och $N(0, \sigma^2)$-fördelade feltermer. Här tolkas interceptet α som den genomsnittliga blodtryckssänkningen i patientgruppen, medan effektparametern β anger den förväntade ytterligare blodtryckssänkningen för en patient som ökar dosen av läkemedlet med 1 mm Hg per dag under en vecka. Resultatet av den kliniska studien sammanfattas med följande summor;

$$\begin{aligned} \textstyle\sum_{i=1}^{20} x_i &= 30.0, \\ \textstyle\sum_{i=1}^{20} (x_i - \bar{x})^2 &= 45.0, \\ \textstyle\sum_{i=1}^{20} Y_i &= 210.0, \\ \textstyle\sum_{i=1}^{20} Y_i(x_i - \bar{x}) &= 202.5, \\ \textstyle\sum_{i=1}^{20} (Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2 &= 105.0, \end{aligned}$$

där $\hat{\alpha}$ och $\hat{\beta}$ är minsta kvadratskattningarna av α och β.

a) Beräkna $\hat{\alpha}$ och $\hat{\beta}$. (3 p)

b) Kalle blir ordinerad av sin husläkare att under en vecka ta 1 g mer per dag av läkemedlet, än genomsnittet i patientgruppen. Bestäm medelfelet $d = \widehat{\mathrm{Var}}(\hat{\mu})^{1/2}$ för skattningen $\hat{\mu}$ av Kalles förväntade blodtryckssänkning $\mu = \alpha + \beta$. (Ledning: Börja med att bestämma $\mathrm{Var}(\hat{\mu})$.) (4 p)

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

Uppgift 2

Vid en kemisk industri framställs en viss typ av plastkassar. Av miljöskäl är det viktigt att plastens hållfasthet är hög, så att en liten mängd plast används till varje kasse. Man vill ta reda på hur känslig plastens hållbarhet är för små variationer, kring nuvarande värden, av de tre viktigaste ingående ämnena. Totalt genomförs $N = 30$ experiment $i = 1, \ldots, 30$ där hållfastheten Y_i och koncentrationerna x_{1i}, x_{2i}, x_{3i} av de tre ämnena registreras. Den fullständiga multipla linjära regressionsmodellen beskrivs av

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

där ε_i är oberoende och $N(0, \sigma^2)$-fördelade feltermer. Dessutom är försöket upplagt så att effekterna av de olika förklarande variablerna är ortogonala, det vill säga

$$\sum_{i=1}^{30} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) = 0$$

för alla $1 \leq j < k \leq 3$, med $\bar{x}_j = \sum_i x_{ji}/30$. För den fullständiga modellen (2) får man följande variansanalystabell:

Variationskälla	Kvs
Ämne 1	8.0
Ämne 2	6.0
Ämne 3	3.0
Residual	23.0
Totalt	40.0

a) För att undersöka om plastens hållfasthet är känslig för variationer i alla tre ämnenas koncentrationer vill kemiingenjörerna på företaget testa olika delmodeller av den fullständiga modellen (2), där en delmängd av de tre ämnena ingår som förklarande variabler. Genomför första steget i bakåteliminering (Backward Elimination, BE). Undersök alltså om någon förklarande variabel ska tas bort från den fullständiga modellen. Signifikansnivån väljs till 5%. (Ledning: På grund av ortogonaliteten mellan de förklarande variablerna fås kvadratsumman för avvikelsen mellan en grund- och en hypotesmodell som summan av kvadratsummorna (i tabellen ovan) för de förklarande variabler som ingår i grundmodellen men inte i hypotesmodellen.) (5 p)

b) Stannar BE-schemat efter a)? Motivera ditt svar. (Ledning: För varje delmodell av den fullständiga modellen ovan så inkluderas de förklarande variabler som inte ingår i delmodellen i variationskällan Residual för delmodellen.) (5 p)

Uppgift 3

I ett land undersöktes hur immuniteten mot COVID-19 varierade mellan olika åldersgrupper och geografiska områden. Forskare vid landets folkhälsomyndighet delade in befolkningen i 4 åldergrupper och 3 regioner. I varje region valdes slumpmässigt 3 personer ut per åldergrupp, bland de som testats positivt för COVID-19 och överlevt sjukdomen. Forskarna ansatte en tvåsidig variansanalys typ I

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, i = 1, 2, 3, 4,\ j = 1, 2, 3,\ k = 1, 2, 3, \qquad (3)$$

för koncentrationen antikroppar hos person k som tillhör åldergrupp i och bor i region j. Här anger μ den genomsnittliga koncentrationen av antikroppar, α_i svarar mot effekten av ålder i, β_j betecknar effekten av region j, samt γ_{ij} samspelet mellan åldergrupp i och region j. Slutligen är alla $\varepsilon_{ijk} \sim N(0, \sigma^2)$ oberoende och normalfördelade feltermer.

a) Modellen i (3) innehåller för många regressionsparametrar μ, $\{\alpha_i\}_{i=1}^4$, $\{\beta_j\}_{j=1}^3$ och $\{\gamma_{ij};\, i = 1, 2, 3, 4, j = 1, 2, 3\}$. Ange hur många fritt varierande regressionsparametrar modellen har, och vilka linjära parameterrestriktioner man kan införa för att åstadkomma detta. (3 p)

b) En variansanalys gav följande resultat:

Variationskälla	Kvs
Ålder	13.5
Region	6.0
Samspel	9.5
Inom celler	25.0
Total	54.0

Testa på nivån 5% om det finns ett signifikant samspel mellan ålder och region vad gäller immunsystemets förmåga att bilda antikroppar. (3 p)
c) Bestäm ett tvåsidigt 95% konfidensintervall för σ. Låt din analys bero av svaret i b), dvs låt variationskällan samspel bidra till att skatta feltermernas varians, om den inte är signifikant. (Ledning: Några av dessa kvantiler för χ^2-fördelningen kan vara till hjälp; $\chi^2_{0.975}(24) = 12.40$, $\chi^2_{0.025}(24) = 39.36$, $\chi^2_{0.975}(30) = 16.79$, and $\chi^2_{0.025}(30) = 46.98$.) (4 p)

Uppgift 4

En statistiskt intresserad restaurangchef genomförde ett 2^2-försök, där kundnöjdheten för en viss typ av nötköttsbiffar undersöktes. De två faktorerna stektid T och typ av sås S varierades båda på två nivåer - och +, svarande mot -1 och +1. För faktorn T svarar - och + mot en kortare respektive längre stektid, medan - och + kodar för de två typerna av sås för faktorn S. För var och en de 4 nivåkombinationerna tillfrågades två kunder. Restaurangchefen ansatte en modell

$$Y_{ijk} = \mu + \bar{T} \cdot i + \bar{S} \cdot j + \overline{TS} \cdot ij + \varepsilon_{ijk}$$

för kundnöjdheten (mätt på en kontinuerlig skala) för kund nummer $k \in \{1, 2\}$ av de som fick äta en biff där stektid och sås hölls på nivåerna $i, j \in \{-, +\}$. Vidare antogs feltermerna ε_{ijk} vara oberoende och normalfördelade med väntevärde 0 och standardavvikelse σ. Resultatet av undersökningen framgår av följande tabell:

T	S	Y_{ij1}	Y_{ij2}
-	-	3.7	4.1
+	-	4.9	4.5
-	+	5.3	5.5
+	+	6.5	6.3

a) Beräkna minsta kvadrat-skattningar $\hat{T}$ och $\hat{S}$ av de två huvudeffekterna, samt $\widehat{TS}$ av samspelseffekten. (Ledning: Börja med att beräkna alla cellmedelvärden $\bar{Y}_{ij\cdot}$.) (4 p)
b) Beräkna en väntevärdesriktig skattning av σ^2. (Ledning: Börja med att beräkna alla $(Y_{ij1} - Y_{ij2})^2$ för att bestämma kvadratsumman inom celler.) (3 p)
c) Beräkna ett tvåsidigt 95% konfidensintervall för den parameter $\bar{T}$ som anger hur stektiden påverkar kundnöjdheten. Kan vi dra slutsatsen att stektiden har en signifikant inverkan på kundnöjdheten? (3 p)

Uppgift 5

Låt

$$Y_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \varepsilon_i, \quad i = 1, \ldots, N$$

vara en multipel linjär regressionsmodell med intercept α och effektparametrar β_1 och β_2 för var och en av de två förklarande variablerna x_1 och x_2. Dessa har centrerats kring $\bar{x}_j = \sum_{i=1}^{N} x_{ji}/N$ för $j = 1, 2$. Feltermerna $\varepsilon_i \sim N(0, \sigma^2)$ antas oberoende.

a) Definiera variationsinflationsfaktorn VIF vid skattning av β_1. (3 p)
b) Definiera förklaringsgraden R_1^2 för en enkel linjär regressionsmodell där x_{1i} är responsvariabel och x_{2i} förklarande variabel. (3 p)
c) Visa att

$$\text{VIF} = \frac{1}{1 - R_1^2}.$$

(Ledning: Du kan ha nytta av matrisinversionsformeln

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix}^{-1} = \frac{1}{ac - b^2} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix}.)$$

(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