STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

30 november 2023 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

En matvaruaffär gjorde en marknadsundersökning för att ta reda på hur mycket graden av tillfredsställelse varierade mellan kunder. En inhyrd statistikkonsult intervjuade 20 olika kunder ($i = 1, \ldots, 20$), som alla bodde i närheten av butiken. Varje kund fick ange ett sammanfattande mått Y_i på hur gärna man handlade i butiken (som i sin tur berodde på närhet, tillgänglighet, priser, service, utbud av varor och deras kvalité). Konsulten misstänkte att närhet till butiken hade en avgörande betydelse. Varje person fick därför även ange hur långt ifrån butiken man bodde (x_i), uttryckt i km. Därefter ansattes en enkel linjär regressionsmodell

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

där ε_i antogs vara oberoende och $N(0, \sigma^2)$-fördelade feltermer. Resultatet av undersökningen framgår av följande variansanalystabell:

Variationskälla	Kvs
Residual	19.3
Regression	5.2
Total	24.5

a) Testa på signifikansnivån 5% nollhypotesen $H_0 : \beta = 0$, att närhet till butiken inte påverkar kundnöjdheten. (3 p)
b) Feltermernas standardavvikelse σ är ett mått på hur mycket kundnöjdheten varierar, när effekten av avstånd till butiken eliminerats. Beräkna en skattning av denna parameter. (Ledning: Börja med att beräkna en väntevärdesriktig skattning av σ^2.) (3 p)
c) Konsulten beräknade ett 95% konfidensintervall $I_\sigma = (0.782, 1.531)$ för σ. Ange vilka två kvantiler av en viss χ^2-fördelning som ingår i detta intervall. Använd sedan 1b) och de undre och övre gränserna för I_σ för att beräkna värdena på dessa två kvantiler. (4 p)

Uppgift 2

Ett försäkringsbolag ville undersöka om det fanns några systematiska skillnader mellan de belopp som betalades ut till kunder i två olika regioner på grund av vattenskador i deras hushåll. Man misstänkte att skadereglerarna vid de två regionskontoren hade olika arbetsrutiner, och samlade in data från 12 hushåll i respektive region, där värdena för region 1 anges först. Man ansatte en multipel linjär regressionsmodell

$$Y_i = \begin{cases} \tilde{\alpha}_1 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{3i} - \bar{x}_3) + \varepsilon_i, & i = 1, \ldots, 12, \\ \tilde{\alpha}_2 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{3i} - \bar{x}_3) + \varepsilon_i, & i = 13, \ldots, 24, \end{cases} \tag{1}$$

för de skadebelopp som olika individer i erhållit (enhet 10,000 SEK). Här anger $\tilde{\alpha}_1$ ett intercept för region 1, medan $\tilde{\alpha}_2$ svarar mot ett intercept för region 2. De resterande tre parametrarna β_1, β_2 och β_3 anger inverkan från bonusnivå i kundens hemförsäkring (x_{1i}), marknadsvärde på fastigheten (x_{2i}), respektive omfattningen av vattenskadan (x_{3i}). Denna inverkan antas vara densamma för båda regionerna, och de tre förklarande variablerna är centrerade utifrån hela datasetet ($\bar{x}_j = \sum_{i=1}^{24} x_{ji}/24$). Vidare antas feltermerna ε_i vara oberoende och $N(0, \sigma^2)$-fördelade.

a) Bestäm parametervektorn $\boldsymbol{\theta}$ för grundmodellen (1). Formulera sedan en hypotesmodell $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\lambda}$ att det inte finns några systematiska skillnader mellan skadeprissättningen i de två regionerna ($\tilde{\alpha}_1 = \tilde{\alpha}_2 = \tilde{\alpha}$, där $\tilde{\alpha}$ är det gemensamma interceptet för båda regionerna), genom att ange matrisen $\boldsymbol{B}$ och kolumnvektorn $\boldsymbol{\lambda}$. (Ledning: $\boldsymbol{\theta}$ och $\boldsymbol{\lambda}$ är kolumnvektorer med fem respektive fyra komponenter, vilket implicerar hur många rader och kolumner $\boldsymbol{B}$ har. Observera att $\boldsymbol{B}\boldsymbol{\lambda}$, då $\boldsymbol{\lambda}$ varierar, svarar mot de värden på $\boldsymbol{\theta}$ som är möjliga enligt hypotesmodellen.) (3 p)

b) Testa hypotesmodellen i a) på signifikansnivån 5%, genom att utgå från följande variansanalystabell:

Variationskälla	Kvs
Avvikelse från hypotes	1.8
Residual	22.3
Total	24.1

(4 p)

c) Ge en kort motivering till varför $\hat{\tilde{\alpha}}_1 = \sum_{i=1}^{12} Y_i/12$ och $\hat{\tilde{\alpha}}_2 = \sum_{i=13}^{24} Y_i/12$ i allmänhet *inte* gäller för grundmodellen, medan skattningen av det gemensamma interceptet $\tilde{\alpha}$ för hypotesmodellen ges av $\hat{\tilde{\alpha}} = \sum_{i=1}^{24} Y_i/24$. (Ledning: Undersök vilka intercept som är centrerade genom att titta på ortogonalitet mellan kolumner hos grundmodellens och hypotesmodellens designmatriser.) (3 p)

Uppgift 3

En stationär MA(1)-process $\{X_t\}$ definieras genom

$$X_t = \varepsilon_t - \theta\varepsilon_{t-1} \tag{2}$$

för $t = \ldots, -1, 0, 1, \ldots$, där $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ är oberoende feltermer medan θ är en godtycklig reellvärd parameter.

a) Använd (2) för att beräkna kovariansfunktionen $\gamma_k = \mathrm{Cov}(X_t, X_{t+k})$ för $k = 0, 1, 2, \ldots$. (4 p)

b) Använda 3a) för att beräkna korrelationsfunktionen $\rho_k = \mathrm{Corr}(X_t, X_{t+k})$ för $k = 0, 1, 2, \ldots$. (2 p)

c) Anta att värdena $\boldsymbol{X}_T = (X_1, \ldots, X_T)$ av processen har observerats för något $T \geq 1$. Härled ett uttryck för prediktorn $\hat{X}_{T+k} = E(X_{T+k}|\boldsymbol{X}_T)$, dels för $k = 1$ och $T = 1$, och dels för $k = 2, 3, 4, \ldots$ och allmänt T. (Ledning: Om Y och Z är två stokastiska variabler med $E(Y) = E(Z) = 0$, $\mathrm{Var}(Y) = \mathrm{Var}(Z)$ och $\rho = \mathrm{Corr}(Y, Z)$ så gäller $E(Z|Y) = \rho Y$.) (4 p)

Uppgift 4

Vid ett forskningslabb ville man uppskatta radioaktiviteten hos två ämnen som ingår i renkött. Man antog att köttet inte innehöll några andra radioaktiva ämnen, och ansatte därför en linjär modell

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

för den uppmätta radioaktiviteten hos prov nummer $i = 1, \ldots, 5$ (enhet: 1000 sönderfall per sekund). Genom masspektroskopi kunde man bestämma mängden x_{1i} och x_{2i} av de två radioaktiva ämnena (enhet: gram) för varje

prov i. De två sökta parametrarna β_1 och β_2 var radioaktiviteten hos respektive ämne (enhet: 1000 sönderfall per gram och sekund). Feltermerna $\varepsilon_i \sim N(0, \sigma^2)$ antogs oberoende. Resultatet av analysen framgår av följande tabell:

i	x_{1i}	x_{2i}	Y_i
1	1.3	2.6	1.6
2	0.7	2.0	1.2
3	1.5	2.0	1.5
4	0.9	1.4	1.2
5	2.1	2.9	1.9

a) Bestäm designmatrisen $\boldsymbol{A}$ för modellen. (Ledning: Eftersom modellen saknar intercept har $\boldsymbol{A}$ två kolumner.) (2 p)

b) Beräkna minsta kvadrat-skattningen $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^T$ av parametervektorn $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$, givet följande information: $\sum_i x_{1i}^2 = 9.65$, $\sum_i x_{2i}^2 = 25.13$, $\sum_i x_{1i}x_{2i} = 15.13$, $\sum_i x_{1i}Y_i = 10.24$ och $\sum_i x_{2i}Y_i = 16.75$. (Ledning: Du kan använda formeln

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

för invertering av en 2×2-matris.) (3 p)

c) Beräkna en väntevärdesriktig skattning $\hat{\sigma}^2$ av feltermsvariansen σ^2, givet att Kvs(Residual) = 0.0909. (2 p)

d) Bestäm en konfidensellipsoid för $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ med konfidensgrad 95%. (Ledning: För att testa ett visst värde $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20})^T$ på $\boldsymbol{\beta}$ används en F-kvot $\|\boldsymbol{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2/(k\hat{\sigma}^2)$, där k är antalet regressionsparametrar i modellen (3). Den konfidensellipsoid man till slut får har den skattade parametervektorn $\hat{\boldsymbol{\beta}}$ i 4b) som mittpunkt.) (3 p)

Uppgift 5

Ett trävaruföretag vill utvärdera kvalitén på den fuktighetsmätare som används. Man valde ut 5 brädor slumpmässigt ur ett stort parti och gjorde 4 mätningar på varje bräda. Detta modellerades som ensidig variansanalys av typ II, för fuktigheten

$$Y_{ij} = \mu + \delta_i + \varepsilon_{ij}, \quad i = 1, \ldots, 5, j = 1, \ldots, 4,$$

vid den j:te mätningen av bräda nummer i. Här anger μ den genomsnittliga fuktigheten hos hela träpartiet, medan $\delta_i \sim N(0, \sigma_\delta^2)$ och $\varepsilon_{ij} \sim N(0, \sigma^2)$ är oberoende stokastiska variabler som beskriver variationen i fuktighet mellan olika brädor respektive mellan mätningar. Man ville undersöka om variansskvoten σ^2/σ_δ^2 var tillräckligt liten för att fuktighetsmätaren skulle anses ha

god kvalité i förhållande till den naturliga fuktighetsvariationen i träpartiet. I tabellen nedan anges stickprovsmedelvärdet $\bar{Y}_{i\cdot}$ (kolumn 2) och stickprovsvariansen s_i^2 (kolumn 3) för mätningarna av de olika brädorna:

Bräda	$\bar{Y}_{i\cdot}$	s_i^2
1	3.1	0.10
2	5.2	0.15
3	4.2	0.18
4	4.6	0.13
5	3.9	0.09

a) Använd alla $\bar{Y}_{i\cdot}$, och deras medelvärde, för att beräkna medelkvadratsumman för variationskällan mellan brädor. (3 p)

b) Använd alla s_i^2 för att beräkna medelkvadratsumman för variationskällan inom brädor. (3 p)

c) Beräkna ett ensidigt 95% konfidensintervall för σ^2/σ_δ^2 av formen $(0, a)$. (Ledning: Börja med att visa att kvoten av medelkvadratsummorna i 5a) och 5b) är fördelad som en F-fördelning multiplicerad med en konstant som beror av σ^2/σ_δ^2. Utgå från formler för de två förväntade medelkvadratsummorna E(Mkvs(Mellan brädor)) och E(Mkvs(Inom brädor)) för att hitta denna konstant.) (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