STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

19 augusti 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

En forskargrupp ville undersöka huruvida risken för att utveckla åldersdiabetes påverkas av vilken variant av en viss gen en person har. Man undersökte 30 diabetespatienter genom att för var och en dem registrera blodglukoshalten (enhet: mmol/l) omedelbart efter en måltid, som ätits på fastande mage. Därefter ansatte man en enkel linjär regressionsmodell

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

för blodglukoshalten hos patient i, som hade $x_i \in \{0, 1, 2\}$ kopior av den variant av genen som man misstänkte var riskförhöjande, medan $\bar{x} = \sum_{i=1}^{30} x_i/30$ anger genomsnittligt antal kopior av denna genvariant för alla patienter. Vidare antas feltermerna ε_i vara oberoende och normalfördelade med väntevärde 0 och varians σ^2. Man delade in patienterna i tre grupper beroende på värdet av den förklarande variabeln, och sammanfattade undersökningen i följande tabell:

x	Antal	Medel	Std
0	10	10.5	2.0
1	10	11.2	2.5
2	10	12.0	3.0

Här anger Medel och Std stickprovsmedelvärdet respektive stickprovsstandardavvikelsen av blodglukoskoncentrationen bland de patienter som hade 0, 1 respektive 2 kopior av den aktuella genvarianten. Syftet med undersökningen var att undersöka om β var positiv.

a) Beräkna minsta kvadrat-skattningen $\hat{\beta}$ av β. (3 p)

b) Beräkna $\mathrm{Var}(\hat{\beta})$, uttryckt i σ^2. (2 p)

c) För att skatta σ^2 ville man *inte* använda sig av residualkvadratsumman från regressionsanalysen, eftersom man misstänkte att modellen (1) var något för enkel, så att residualerna kunde fånga upp ett icke-linjärt samband mellan $E(Y_i)$ och x_i. Använd istället stickprovsstandardavvikelserna ovan för att skatta σ^2. Beräkna därefter medelfelet för skattningen $\hat{\beta}$. (3 p)

d) Beräkna ett ensidigt konfidensintervall för β av typ (a, ∞) med konfidensgrad 97.5%. Ange sedan huruvida vi kan dra slutsatsen att β är positiv. (2 p)

Uppgift 2

Ett läkemedelsföretag har utvecklat en njurmedicin för att sänka kreatinhalten i blodet hos patienter med njurproblem. Man ville bestämma hur väl medicinen fungerade genom att mäta kreatinhalten i blodet hos patienter som fick dosen $i = 1, 2, 3, 4, 5$ mg av medicinen. Man ansatte grundmodellen

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, 3, 4, 5,\ j = 1, \ldots, n_i,$$

för kreatinhalten hos patient nummer j inom det stickprov med n_i individer som fick dosen i. Här anger μ_i förväntad kreatinhalt för individer med dos i, medan ε_{ij} är oberoende och normalfördelade feltermer med väntevärde 0 och varians σ^2. Vidare studerade man hypotesmodellen

$$H_0 : \mu_i = \alpha + \beta i$$

att den förväntade kreatinhalten berodde linjärt av dosen av njurmedicinen. Resultatet av studien framgår av följande variansanalystabell:

Variationskälla	Kvs
Linjär regression	6.8
Icke-linjäritet	35.5
Inom stickprov	46.0
Total	88.3

a) Beräkna antalet frihetsgrader för de tre variationskällorna, om $n_1 = n_2 = n_4 = n_5 = 3$ och $n_3 = 8$. (Ledning: Totala antalet frihetsgrader är $N - 1 = \sum_{i=1}^{5} n_i - 1$.) (2 p)

b) Testa på nivån 5% om icke-linjäriteten är signifikant. (4 p)

c) Beräkna minsta kvadrat-skattningen $\hat{\beta}$ av β, om man vet att $\beta < 0$. (Ledning: Börja med att bestämma $\hat{\beta}^2$ utifrån en av kvadratsummorna.) (4 p)

Uppgift 3

Vid ett lantbruksuniversitet undersöktes hur tillväxthastigheten av en viss växt (enhet: g/dag) påverkades av tre faktorer; kvävehalten i jorden K, bevattningsmängden V, samt belysningsgraden B. Man utgick från standardnivåer på dessa tre faktorer, svarande mot en förväntad tillväxthastighet μ. Tidigare försök hade indikerat att standardnivåerna gav den optimala förväntade tillväxthastigheten. För att undersöka om så var fallet genomfördes ett 2^3-försök utan replikat, där man varierade alla tre faktorer på en hög (+) och en låg (-) nivå, på samma avstånd över och under respektive standardnivå. Man ansatte en modell

$$Y_{ijk} = \mu + \bar{K} \cdot i + \bar{V} \cdot j + \bar{B} \cdot k + +\overline{KV} \cdot ij + \varepsilon_{ijk},$$

för tillväxthastigheten då de tre faktorerna ligger på nivåerna $i, j, k \in \{-, +\}$, svarande mot -1 och +1. Vidare anger $\bar{K}$, $\bar{V}$ och $\bar{B}$ huvudeffekterna av respektive faktor, dvs den förväntade effekten av att höja gödningsmängden kväve, bevattningsgraden respektive ljusmängden, $\overline{KV}$ anger samspelet mellan kväve och vatten, medan $\varepsilon_{ijk} \sim N(0, \sigma^2)$ är oberoende feltermer. Resultatet av datainsamlingen framgår av följande tabell:

K	V	B	Y_{ijk}
-	-	-	10.0
+	-	-	11.0
-	+	-	11.5
+	+	-	12.2
-	-	+	9.5
+	-	+	10.0
-	+	+	10.5
+	+	+	11.0

a) Beräkna minsta kvadrat-skattningar $\hat{K}$, $\hat{V}$ och $\hat{B}$ av de tre huvudeffekterna och $\widehat{KV}$ av samspelseffekten mellan K och V. (3 p)

b) Alla fyra skattningarna i a) har samma varians. Beräkna denna varians, uttryckt i σ^2. (3 p)

c) Beräkna en skattning av σ^2, genom att använda skattningarna i a) samt att $\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2 = 5.529$. Använd detta för att dra slutsatsen vilka

konfidensintervall med konfidensgrad 95%, för $\bar{K}$, $\bar{V}$, $\bar{B}$ respektive $\overline{KV}$, som täcker över 0. Kan man säga att standardnivåerna för de tre faktorerna ger optimal förväntad tillväxthastighet för växten? (Ledning: Du kan utan bevis utnyttja att för var och en av de fyra faktorerna $\theta \in \{K, V, B, KV\}$ gäller att Kvs(Faktor θ) $= 8\hat{\theta}^2$.) (4 p)

Uppgift 4

Vid en stålindustri försökte man maximera hårdheten hos en viss metallegering. Speciellt undersökte man om en ny metall M skulle ingå i legeringen, genom att vart och ett av 6 provbad (som innehöll de andra metallerna, i smält form) delades upp i två delar. Till den ena delen av varje provbad tillsättes en på förhand bestämd proportion av M, medan den andra delen inte fick någon tillsats av M. Efter stelning uppmätte man hårdheten hos de erhållna legeringarna. Man ställde upp den blandade modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, 2,\ j = 1, \ldots, 6,$$

för hårdheten hos den legering som erhölls från provbad j, utan ($i = 1$) eller med ($i = 2$) tillsats av M. Effekten av tillsats antogs vara en systematisk faktor med $\alpha_2 = -\alpha_1$, medan provbad betraktades som en slumpmässig faktor, med oberoende $\beta_j \sim N(0, \sigma_\beta^2)$, svarande mot att proportionerna av de övriga metallerna varierade något mellan provbad. Vidare antogs ε_{ij} vara oberoende och normalfördelade feltermer med väntevärde 0 och varians σ^2. Resultatet av experimenten sammanfattades i följande variansanalystabell:

Variationskälla	Kvs
Tillsats av M	3.3
Provbad	10.4
Residual	5.2
Total	18.9

a) Man var primärt intresserade att att skatta effekten $\Delta = \alpha_2 - \alpha_1 = 2\alpha_2$ av att tillsätta M. Bestäm minsta kvadrat-skattningen av $\hat{\Delta}$ av Δ, om vi vet att $\hat{\Delta} > 0$. (Ledning: $\hat{\Delta}^2$ kan bestämmas utifrån Kvs(Tillsats av M).) (3 p)

b) Beräkna en skattning av σ^2. (3 p)

c) Bestäm medelfelet hos skattningen $\hat{\Delta}$, och ange därefter ett 95% konfidensintervall för Δ. (Ledning: Börja med att beräkna Var($\hat{\Delta}$).) (4 p)

Uppgift 5

En multipel linjär regressionsmodell med två kovariater och utan intercept, kan skrivas

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

där Y_i är responsvariabelns värde för observation i, x_{i1} och x_{i2} de två kovariaternas värden för observation i, β_1 och β_2 de två effektparametrarna för respektive kovariat, samt $\varepsilon_i \sim N(0, \sigma^2)$ oberoende feltermer.

a) Skriv modellen på matrisform

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

där $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ är en matris med två kolumner $\boldsymbol{x}_1$ och $\boldsymbol{x}_2$. Definiera $\boldsymbol{Y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ och $\boldsymbol{\varepsilon}$. (1 p)

b) Definiera variationsinflationsfaktorn (VIF) vid skattning av β_1. Visa att

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

om $c = \boldsymbol{x}_1^T \boldsymbol{x}_2 / \sqrt{\boldsymbol{x}_1^T \boldsymbol{x}_1 \cdot \boldsymbol{x}_2^T \boldsymbol{x}_2}$ är korrelationskoefficienten mellan de två kolumnerna i designmatrisen $\boldsymbol{X}$ ($-1 < c < 1$). Kommentera resultatet i termer av kolinearitet. (Ledning: Du kan ha användning av formeln

$$\begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}^{-1} = \frac{1}{\alpha\beta - \gamma^2} \begin{pmatrix} \beta & -\gamma \\ -\gamma & \alpha \end{pmatrix}$$

för invertering av kvadratiska, symmetriska matriser av ordning 2.) (3 p)

c) Definiera förklaringsgraden R^2 för modellen (2), samt förklaringsgraden R_1^2 och R_2^2 för de två delmodeller av (2) som bara tar med kovariat 1 respektive kovariat 2. Visa sedan att

$$R^2 = \frac{1}{1 - c^2} \left(R_1^2 + R_2^2 - 2cR_1R_2 \right), \tag{3}$$

där $R_j = \sqrt{R_j^2}$, om vi vet att minsta kvadrat-skattningen $\tilde{\beta}_1$ av β_1, för delmodellen med bara kovariat 1, och minsta kvadrat-skattningen $\tilde{\beta}_2$ av β_2, för delmodellen med bara kovariat 2, båda är positiva. (Ledning: Eftersom modellen (2) inte har med intercept så vill vi inte förklara variationen av Y_i kring $\bar{Y}_\cdot$, utan kring 0. Ersätt därför $\bar{Y}_\cdot$ med 0 överallt i den vanliga definitionen av förklaringsgrad.) (4 p)

d) Man mätte upp strålningshalten Y_i från N provbitar, som alla innehöll samma radioaktiva ämne. Man var intresserad av strålningshalten per viktenhet av detta ämne, och antog vidare att ingen av proverna innehöll någon annan källa till radioaktivitet. Eftersom det var svårt att viktbestämma det radioaktiva ämnet använde man två olika mätmetoder, så att x_{1i} och x_{2i} var de uppmätta vikten för prov i, med metod 1 respektive metod 2. Vid analysen anpassade man dels den fulla modellen (2) till data, samt dels de båda delmodellerna där endast kovariat 1 respektive kovariat 2 togs med. För den första delmodellen erhölls en effektskattning $\tilde{\beta}_1 > 0$ och för den andra delmodellen $\tilde{\beta}_2 > 0$. Vidare räknade man ut följande förklaringsgrader:

Modell	Förklaringsgrad
Kovariat 1	$R_1^2 = 0.3$
Kovariat 2	$R_2^2 = 0.4$
Kovariat 1 och 2	$R^2 = 0.7$

Dessa värden stämmer enligt (3) med att de två viktmätningsmetoderna var ortogonala ($c = 0$). I detta fall visste man dock att de två mätmetoderna var (starkt) positivt korrelerade men inte identiska ($0 < c < 1$ nära 1). Visa att det finns ett sådant värde på c som stämmer med förklaringsgraderna i tabellen, och bestäm motsvarande variationsinflationsfaktor. (Ledning: Du kan lösa d) genom att använda b) och c), utan att ha löst dessa deluppgifter.) (2 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