STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

20 oktober 2022 14–19

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

För att undersöka prisutvecklingen hos en viss typ av päron uppskattades det genomsnittliga kilopriset y_i i Sverige under nio på varandra följande år ($x_1 = 2012, x_2 = 2013, \ldots, x_9 = 2020$). Man ansatte en enkel linjär regressionsmodell

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, 9,$$

där y_i är en observation av Y_i, $\bar{x} = \sum_{i=1}^{9} x_i/9$ och $\varepsilon_i \sim N(0, \sigma^2)$ är oberoende feltermer.

a) Beräkna en skattning av det förväntade kilopriset år 2022 om $\bar{y} = \sum_{i=1}^{9} y_i/9 = 26.9$ och $\sum_{i=1}^{9} y_i(x_i - \bar{x}) = 27.3$. (4 p)

b) Beräkna medelfelet d för skattningen i a), om kvadratsumman för variationskällan "Residual" i variansanalystabellen ges av Kvs = 0.9274. (3 p)

c) Ange ett 95% konfidensintervall för det förväntade kilopriset år 2022. (Ledning: Intervallet innehåller kvantilen $t_{0.025}(7) = 2.3646$.) (3 p)

Uppgift 2

Ett läkemedelsföretag har sedan länge astmamedicinen 1 på marknaden och vill undersöka om en annan, nyutvecklad medicin 2 har samma effekt. För respektive medicin uppmättes syreupptagningsförmågan Y hos 5 patienter en timme efter att de tagit en dos (olika för de 5 patienterna). Sedan antogs en linjär dos respons-kurva för respektive medicin. Det svarar mot en linjär modell

$$Y_{ij} = \alpha_i + \beta_i x_j + \varepsilon_{ij}, \quad i = 1, 2, \; j = 1, \ldots, 5, \tag{1}$$

för syreupptagningsförmågan hos den patient som tagit dosen x_j mg av medicin i, där feltermerna $\varepsilon_{ij} \sim N(0, \sigma^2)$ antas vara oberoende.

a) Ange observationsvektor $\boldsymbol{Y}$, designmatris $\boldsymbol{A}$, parametervektor $\boldsymbol{\theta}$ och feltermsvektor $\boldsymbol{\varepsilon}$ för den linjära modellen

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

som svarar mot (1). (Ledning: $\boldsymbol{\theta}$ har fyra element, två intercept och två effektparametrar.) (3 p)

b) Nollhypotesen

$$H_0 : \alpha_1 = \alpha_2, \; \beta_1 = \beta_2$$

att de två medicinerna har samma effekt kan formuleras som $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\lambda}$ för lämpligt valda $\boldsymbol{B}$ och $\boldsymbol{\lambda}$. Ange $\boldsymbol{B}$ och $\boldsymbol{\lambda}$. (Ledning: $\boldsymbol{\lambda}$ har två element.) (3 p)

c) Nedan anges ett utdrag ur variansanalystabellen för försöket. Använd det för att testa nollhypotesen på nivån 5%.

Variationskälla	Kvs
Avvikelse från H_0	6.8910
Residual	1.1060
Totalt	7.9970

(4 p)

Uppgift 3

Ett byggföretag utför en kvalitetskontroll genom att mäta fuktigheten i ett betongparti som består av ett antal olika betongblock. Man väljer ut ett stickprov på 5 block och genomför 3 mätningar på varje block. De uppmätta fuktigheterna Y_{i1}, Y_{i2}, Y_{i3} för block $i = 1, \ldots, 5$ antas följa en ensidig variansanalysmodell

$$Y_{ij} = \mu + \delta_i + \varepsilon_{ij}$$

av typ II, med väntevärde μ, oberoende blockeffekter $\delta_i \sim N(0, \sigma_\delta^2)$ och mätfel $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. Resultatet sammanfattas i följande tabell, i form av stickprovsmedelvärden $\bar{Y}_{i\cdot}$ och stickprovsvarianser s_i^2 för mätningarna inom respektive betongblock:

Block	$\bar{Y}_{i\cdot}$	s_i^2
1	3.1	0.035
2	2.6	0.031
3	2.8	0.020
4	3.3	0.046
5	2.7	0.023

a) Skatta de två varianskomponenterna σ_δ^2 och σ_ε^2. (Ledning: Du har nytta av att beräkna stickprovsvariansen av $\bar{Y}_{i\cdot}$, $i = 1, \ldots, 5$.) (5 p)
b) Ange ett 95% konfidensintervall för μ. (Ledning: Den t-kvantil du behöver fås som kvadratroten ur en lämplig F-kvantil i tabellen nedan.) (5 p)

Uppgift 4

Låt $y_1, \ldots, y_T$ vara observationer från en AR(1)-process

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t, \tag{2}$$

där $\mu = E(Y_t)$ och $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ är oberoende och normalfördelade.

a) För vilka värden på ϕ är processen stationär? (Inget bevis krävs.) (1 p)

b) Givet värden på ϕ enligt a), härled värden på $\sigma^2 = \mathrm{Var}(Y_t)$ och autokorrelationsfunktionen $\rho_k = \mathrm{Corr}(Y_t, Y_{t+k})$ för $k = 0, 1, 2, \ldots$. (4 p)

c) Parametervektorn för en AR(1)-process kan skrivas $\boldsymbol{\psi} = (\mu, \phi, \sigma_\varepsilon^2)$. Låt $f(y_t|y_{t-1})$ ange den betingade täthetsfunktionen för Y_t given $Y_{t-1} = y_{t-1}$ och definiera en approximativ log likelihood

$$\begin{aligned} l(\boldsymbol{\psi}) &= \sum_{t=2}^T \log f(y_t|y_{t-1}) \\ &= -\sum_{t=2}^T \left[\log(\sqrt{2\pi}\sigma_\varepsilon) + \frac{(y_t-\mu-\phi(y_{t-1}-\mu))^2}{2\sigma_\varepsilon^2}\right]. \end{aligned} \tag{3}$$

Bestäm ML-skattningarna $\hat{\mu}$ och $\hat{\phi}$ av μ och ϕ, baserat på den approximativa log likelihooden (3). (Ledning: Skriv om (2) som $Y_t = \alpha + \phi Y_{t-1} + \varepsilon_t$, där $\alpha = \mu(1-\phi)$. Börja med att bestämma ML-skattningar $\hat{\alpha}$ och $\hat{\phi}$ av α och ϕ, som inte kommer att bero på värdet av σ_ε^2. Ditt svar kan uttryckas med hjälp av bland annat $\bar{x} = \sum_{t=1}^{T-1} y_t/(T-1)$ och $\bar{y} = \sum_{t=2}^T y_t/(T-1)$.) (5 p)

Uppgift 5

Anta att vi har en multipel linjär regressionsmodell

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

med observationsvektor $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$, designmatris $\boldsymbol{A}$, regressionsparametrar $\boldsymbol{\theta} = (\alpha, \beta_1, \ldots, \beta_{k-1})^T$ och feltermsvektor $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_N)$, där $\boldsymbol{I}_N$ är en identitetsmatris av ordning N. Minsta kvadrat-skattningen $\hat{\boldsymbol{\mu}} =$

$(\hat{\mu}_1, \dots, \hat{\mu}_N)^T$ av väntevärdesvektorn $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\theta}$ kan skrivas med hjälp av hattmatrisen $\boldsymbol{H} = (h_{ij})_{i,j=1}^N$ som

$$\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{Y}. \tag{4}$$

a) Utgå från minsta kvadrat-skattningen av $\boldsymbol{\theta}$ och visa att hattmatrisen kan uttryckas endast med hjälp av designmatrisen enligt

$$\boldsymbol{H} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T.$$

(3 p)

b) Låt $e_i = Y_i - \hat{\mu}_i$ vara residualen för observation i. Visa att

$$E(e_i^2) = \sigma^2(1 - h_{ii}). \tag{5}$$

(Ledning: Det kan vara enklast att först undersöka kovariansmatrisen $\mathrm{Var}(\boldsymbol{e}) = \mathrm{Var}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$ för hela residualvektorn $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}} = (e_1, \dots, e_N)^T$, genom att utnyttja (4) och egenskaper hos $\boldsymbol{H}$.) (3 p)

c) Härled hattmatrisen för enkel linjär regression, där de förklarande variablerna före centrering har värdena $x_1, \dots, x_N$. Använd sedan (5) för att först ge ett uttryck för $E(e_i^2)$, och sedan diskutera vilka observationspunkter (x_i, Y_i) som kan förväntas vara inflytelserika. (Ledning: Börja med att ställa upp designmatrisen. När du gör det får du enklast räkningar om du först centrerar de förklarande variablerna i regressionsmodellen.) (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