STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

25 oktober 2024 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

Den statistikintresserade elitmotionären Lasse är en före detta tiokampare med följande löprekord på 6 olika löpsträckor:

Sträcka i	Distans d_i	Löptid T_i
1	100 m	11.8 s
2	200 m	25.3 s
3	400 m	53.1 s
4	800 m	2 min 03 s
5	1500 m	3 min 58 s
6	5000 m	14 min 10 s

För att undersöka hur hans löprekord påverkas av distans så logaritmerar Lasse först löpsträckorna ($x_i = \log(d_i)$) och löptiderna ($y_i = \log(T_i)$), med d_i angiven i meter och T_i i sekunder. Därefter ansätter han den enkla linjära regressionsmodellen

$$Y_i = \tilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i = \mu(x_i) + \varepsilon_i, \quad i = 1, \ldots, 6, \tag{1}$$

för de logaritmerade löptiderna, där y_i antas vara en observation av den stokastiska variabeln Y_i. Vidare är $\varepsilon_1, \ldots, \varepsilon_6$ oberoende och normalfördelade feltermer med väntevärde 0 och varians σ^2, medan $\bar{x} = \sum_{i=1}^{6} x_i/6$ är medelvärdet av alla x_i. Med ett statistikprogram räknar Lasse ut att

$$\begin{aligned} \bar{x} &= 6.4017, \\ \textstyle\sum_{i=1}^{6} y_i &= 26.7008, \\ \textstyle\sum_{i=1}^{6} (x_i - \bar{x})^2 &= 9.9995, \\ \textstyle\sum_{i=1}^{6} (x_i - \bar{x}) y_i &= 10.9916. \end{aligned}$$

a) Beräkna minsta kvadratskattningarna av interceptet $\tilde{\alpha}$ och av lutningsparametern β i regressionsmodellen (1). (3 p)

b) Ange kovariansmatrisen $\mathrm{Var}(\hat{\tilde{\alpha}}, \hat{\beta})$ för de två skattningarna i a), uttryckt i feltermsvariansen σ^2. (2 p)

c) Från en variansanalystabell med variationskällorna Regression och Residual får Lasse fram att Kvs(Residual) = 0.004757. Använd detta för att skatta σ^2. (2 p)

d) Lasses förväntade löptid på en engelsk mile (1609 m) är $t = \exp(\mu(x_0))$, där $x_0 = \log(1609) = 7.3834$. Beräkna ett 95% konfidensintervall för t. (Ledning: Börja med att räkna ut $\mathrm{Var}[\hat{\mu}(x_0)]$ och ett 95% konfidensintervall för $\mu(x_0)$, med hjälp av resultaten från a)-c). Du har då nytta av att $t_{0.025}(4) = \sqrt{F_{0.05}(1,4)}$.) (3 p)

Uppgift 2

En balansvåg med två skålar A och B mäter differensen i vikt mellan det som placerats i skål A och skål B, plus ett systematiskt fel α, och ett slumpfel, som antas normalfördelat med väntevärde 0 och varians σ^2. Dessutom är slumpfelen oberoende mellan mätningar. För att uppskatta vikterna β_1 och β_2 av två objekt 1 och 2, så genomfördes följande fyra mätningar (enhet: kg):

Mätning i	Objekt i skål A	Objekt i skål B	Vågens utslag y_i
1	1 och 2	Inga	5.2
2	Inga	1 och 2	-4.8
3	1	2	-0.6
4	2	1	0.8

Här anger alltså ett positivt (negativt) värde på y_i att skål A (B) väger över jämfört med den andra skålen i respektive mätning.

a) De fyra viktmätningarna kan formuleras som en multipel linjär regressionsmodell $\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon}$, med responsvektor $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^T$, där y_i är en observation av Y_i, designmatris $\boldsymbol{A}$, regressionsparametrar $\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2)^T$, och feltermsvektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)^T$. Ange designmatrisen. (2 p)

b) Beräkna minsta-kvadratskattningarna av α, β_1 och β_2. (3 p)

c) Beräkna variansen för skattningen av β_1 i b), uttryckt i σ^2. (2 p)

d) Räkna ut ett 95% konfidensintervall för vikten β_1 av objekt 1. (Ledning: För att räkna ut medelfelet, börja med att bestämma residualer, där du utan bevis kan använda att minsta kvadrat-skattningen av $\boldsymbol{\mu} = \boldsymbol{A\theta}$ är $\hat{\boldsymbol{\mu}} = \boldsymbol{A}\hat{\boldsymbol{\theta}} = (5.15, -4.85, -0.55, 0.85)^T$. Beräkna därefter en väntevärdesriktig skattning av feltermernas varians σ^2.) (2 p)

e) Om σ hade varit känd, vilken hade den simultana konfidensgraden för det 95% konfidensintervallet för β_1 *och* motsvarande 95% konfidensintervall för vikten β_2 av objekt 2 varit? (Ledning: Uppgiften kräver inga långa uträkningar, och inga nya konfidensintervall behöver räknas ut.) (1 p)

Uppgift 3

En forskargrupp undersökte hur ett företags storlek och bransch påverkade de anställdas tillfredsställelse eller trivsel med sitt arbete. Man delade in företagen i tre kategorier efter storlek (1=liten, 2=mellanstor, 3=stor) och även i 5 kategorier efter bransch. För alla $3 \times 5 = 15$ kombinationer av storlek och bransch valde man slumpmässigt ut 2 företag. Sedan intervjuades slumpmässigt en person från var och en av alla $3 \times 5 \times 2 = 30$ utvalda företag. Låt Y_{ijk} beteckna trivseln hos den person som intervjuades från företag $k = 1, 2$, bland de som tillhör storlekskategorin i från branschkategorin j. Forskarna ansatte en tvåsidig variansanalys typ I, dvs

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, i = 1, 2, 3,\ j = 1, 2, 3, 4, 5,\ k = 1, 2, \quad (2)$$

där μ anger de anställdas genomsnittliga trivsel bland de undersökta företagen, α_i effekten av storlek i, β_j effekten av bransch j, samt γ_{ij} samspelet mellan storlek i och bransch j. Slutligen är alla $\varepsilon_{ijk} \sim N(0, \sigma^2)$ oberoende och normalfördelade feltermer.

a) Modellen i (2) innehåller för många regressionsparametrar μ, $\{\alpha_i\}_{i=1}^3$, $\{\beta_j\}_{j=1}^5$ och $\{\gamma_{ij};\, i = 1, 2, 3, j = 1, \ldots, 5\}$. Ange hur många fritt varierande regressionsparametrar modellen bör ha, och vilka linjära parameterrestriktioner man kan införa för att åstadkomma detta. (3 p)

b) En variansanalys gav följande resultat:

Variationskälla	Kvs
Storlek	22.60
Bransch	31.40
Samspel	24.80
Inom celler	30.75
Total	109.55

Testa på nivån 5% om det finns ett signifikant samspel mellan storlek och bransch vad gäller de anställdas tillfredsställelse. (3 p)

c) Testa på nivån 5% om företagets storlek har en signifikant inverkan på de anställdas trivsel. Låt din analys bero av svaret i b), dvs låt variationskällan samspel bidra till att skatta feltermernas varians, om den inte är signifikant. (4 p)

Uppgift 4

Låt $\{Y_t\}$ vara en stationär AR(2)-process, det vill säga $X_t = Y_t - E(Y_t) = Y_t - \mu$ ges av

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t, \tag{3}$$

med oberoende feltermer $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, medan ϕ_1 och ϕ_2 är de två autoregressionsparametrarna.

a) För vilka värden på ϕ_1 och ϕ_2 är processen stationär? (Ledning: Svaret kommer att involvera rötter till en viss andragradsekvation.) (2 p)
b) Låt $\rho_k = \mathrm{Corr}(Y_t, Y_{t+k}) = \mathrm{Corr}(X_t, X_{t+k})$, $k = 0, 1, 2, \ldots$ beteckna autokorrelationsfunktionen. Visa att

$$\begin{aligned} \rho_1 &= \phi_1/(1-\phi_2), \\ \rho_2 &= \phi_2 + \phi_1^2/(1-\phi_2). \end{aligned}$$

(Ledning: Börja med att utnyttja (3) för att ge uttryck för $\gamma_1 = \mathrm{Cov}(X_t, X_{t+1})$ och $\gamma_2 = \mathrm{Cov}(X_t, X_{t+2})$ som innefattar $\gamma_0 = \mathrm{Var}(X_t)$. Uttryck därefter ρ_k med hjälp av γ_k och γ_0.) (5 p)
c) Utnyttja 4a) för att visa att $|\phi_2| < 1$ alltid gäller för en stationär AR(2)-process. Ge vidare exempel på en stationär AR(2)-process med $|\phi_1| > 1$. (3 p)

Uppgift 5

Låt

$$Y_i = \sum_{j=1}^{m} \beta_j x_{ji} + \varepsilon_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{4}$$

vara *träningsdata* för en multipel linjär regressionsmodell utan intercept med m förklarande variabler, som antar värdena $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{mi})^T$ för observation i, och effektparametrar $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T$. Vidare antas feltermerna $\varepsilon_i \sim N(0, \sigma^2)$ vara oberoende och normalfördelade. Man vill undersöka modellens prediktionsförmåga med korsvalideringskriteriet

$$\mathrm{CV} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{\boldsymbol{\beta}}_{(-i)}^T \boldsymbol{x}_i)^2, \tag{5}$$

där $\hat{\boldsymbol{\beta}}_{(-i)}$ är minsta kvadrat-skattningen av $\boldsymbol{\beta}$ då alla $N-1$ observationer i träningsdata utom i tagits med. Även om CV endast involverar träningsdata (4), så kan det ses som en skattning av det genomsnittliga kvadratiska

prediktionsfelet för ett nytt *testdataset* med samma värden på de förklarande variablerna som i (4).

a) Notera att $Y_i - \hat{\boldsymbol{\beta}}_{(-i)}^T \boldsymbol{x}_i$ är ett *prediktionsfel*, då Y_i predikteras med hjälp av de andra observationerna i träningsdata (4). Använd (4) för att uttrycka detta prediktionsfel med hjälp av feltermen ε_i för observation i, $\boldsymbol{x}_i$ och *skattningsfelet* $\hat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}$ av effektparametrarna då observation i ej tagits med. (2 p)

b) Skattningsfelet i a) har väntevärde $(0, \ldots, 0)^T$ och $p \times p$ kovariansmatris $\boldsymbol{C}_i = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{(-i)})$. Ange ett uttryck för $E(\mathrm{CV})$ som innehåller σ^2, samt $\boldsymbol{x}_i$ och $\boldsymbol{C}_i$ för $i = 1, \ldots, N$. (Ledning: Utnyttja a), räkneregel för $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{(-i)}^T \boldsymbol{x}_i)$ samt att ε_i och $\hat{\boldsymbol{\beta}}_{(-i)}$ är oberoende stokastiska variabler. Du behöver inte ange ett uttryck för $\boldsymbol{C}_i$.) (3 p)

c) Observera att de m förklarande variablerna inte är centrerade och ange därefter designmatrisen $\boldsymbol{X}$ för modellen. Låt $\hat{\boldsymbol{\beta}}$ vara minsta kvadrat-skattningen av $\boldsymbol{\beta}$ baserat på alla N observationer. Uttryck sedan $\boldsymbol{C} = \mathrm{Var}(\hat{\boldsymbol{\beta}})$ med hjälp av σ^2 och $\boldsymbol{X}$. (2 p)

d) Om N är stor kan vi vidare anta att $\boldsymbol{C}_i \approx \boldsymbol{C}$. Uttnyttja detta, och resultaten i b) och c), för att visa att

$$E(\mathrm{CV}) \approx \sigma^2(1 + \frac{m}{N}). \tag{6}$$

(Ledning: När du ersatt $\boldsymbol{C}_i$ med $\boldsymbol{C}$, får du efter förenkling ett uttryck som innehåller element från hattmatrisen $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ för regressionsmodellen. Du får utan bevis anta att summan av diagonalelementen i hattmatrisen är m.) (2 p)

e) Ge en tolkning av den andra termen $\sigma^2 m/N$ i högerledet av (6). (Ledning: Börja med att visa att den andra termen i högerledet i (6) försvinner när $\boldsymbol{\beta}$ är känd och alltså $\hat{\boldsymbol{\beta}}_{(-i)}$ kan ersättas med $\boldsymbol{\beta}$ i (5). Uppgiften kräver inga långa räkningar.) (1 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