STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Tentamen för kursen

Linjära statistiska modeller

18 augusti 2023 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

Betrakta den enkla linjära regressionsmodellen med responsvariabler

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

och oberoende feltermer ε_i som uppfyller $E(\varepsilon_i) = 0$ och $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Låt vidare $\hat{\tilde{\alpha}}$ och $\hat{\beta}$ vara minsta kvadrat-skattningarna av $\tilde{\alpha}$ och β.

a) Uttryck kovariansmatrisen

$$\begin{pmatrix} \mathrm{Var}(\hat{\tilde{\alpha}}) & \mathrm{Cov}(\hat{\tilde{\alpha}}, \hat{\beta}) \\ \mathrm{Cov}(\hat{\tilde{\alpha}}, \hat{\beta}) & \mathrm{Var}(\hat{\beta}) \end{pmatrix}$$

för parameterskattningarna $\hat{\tilde{\alpha}}$ och $\hat{\beta}$ med hjälp av feltermsvariansen σ^2, de förklarande variablerna x_i och antalet observationer N. (3 p)

b) Ange ytterligare en egenskap hos feltermerna $\{\varepsilon_i\}_{i=1}^N$ som medför att skattningarna $\hat{\tilde{\alpha}}$ och $\hat{\beta}$ av intercept och lutning är oberoende stokastiska variabler. (2 p)

c) Låt x_i och Y_i svara mot inflationstakten och bostadsräntan (båda uttryckta i procent) under N olika tidpunkter, samt anta att Y_i är normalfördelad. Parametern β anger alltså hur känslig räntan är för ändringar i inflationstakten. Man utgick från ett datamaterial där inflationstakt och räntesats observerades vid $N = 20$ tidpunkter med så stort avstånd att oberoendeantagandet mellan feltermerna i (1) kan antas gälla. Ange ett 95% konfidensintervall för β om minsta kvadrat-skattningen är $\hat{\beta} = 0.8$, medelkvadratsumman för residualer är Mkvs(Residual) $= 0.45$, samt $\sum_{i=1}^{20}(x_i - \bar{x})^2 = 29.2$. (Ledning: Du kan använda den bifogade tabellen med F-kvantiler för att beräkna den t-kvantil du behöver genom sambandet $t_{p/2}(f) = \sqrt{F_p(1, f)}$.) (5 p)

Uppgift 2

Ett företag är stationerat i $N = 20$ olika regioner. Man undersöker hur den sålda kvantiteten Y_i av dammsugare under ett år i de olika regionerna $i = 1, \ldots, 20$ förklaras av fyra variabler enligt en multipel linjär regressionsmodell (grundmodell). Dessa förklarande variabler är pris (x_{1i}), antal konkurrerande företag (x_{2i}), hushållens medelinkomst (x_{3i}) och genomsnittliga bostadsyta (x_{4i}) i respektive region. En minsta kvadrat-analys gav förklaringsgraden $R_0^2 = 0.756$.

a) Uttryck R_0^2 med hjälp av responsvariablerna Y_i, deras medelvärde $\bar{Y}$ och skattade väntevärden $\hat{\mu}_i$. (3 p)

b) Eftersom medelinkomst och bostadsyta är korrelerade gjordes även en anpassning till en hypotesmodell utan bostadsyta som förklarande variabel. Men erhöll då förklaringsgraden $R_1^2 = 0.721$. Uttryck $R_0^2 - R_1^2$ med hjälp av responsvariablerna, deras medelvärde $\bar{Y}$ och skattade väntevärden $\hat{\mu}_i$ och $\hat{\hat{\mu}}_i$ enligt grund- och hypotesmodellen. (Ledning: Börja med att uttrycka R_1^2 som i a), fast med $\hat{\hat{\mu}}_i$ istället för $\hat{\mu}_i$.) (3 p)

c) Utnyttja a) och b) för att visa att

$$\text{F-kvot} = \frac{(R_0^2 - R_1^2)/(k - l)}{(1 - R_0^2)/(N - k)}$$

när grundmodellen (med k parametrar) testas mot hypotesmodellen (med l parametrar). (2 p)

d) Testa på signifikansnivån 5% om grundmodellen ger en signifikant bättre anpassning till data än hypotesmodellen. (Ledning: Börja med att ange k och l. Både grundmodellen och hypotesmodellen innehåller, förutom effektparametrar, ett intercept.) (2 p)

Uppgift 3

En medicinsk forskargrupp misstänker att en muterad variant en viss gen orsakar försämrad ämnesomsättning. Sedan tidigare vet man att ämnesomsättningen även påverkas av graden av motion (miljö- eller livsstilseffekt). För

att undersöka om en genetisk komponent föreligger registrerades ämnesomsättningen Y_{ijk} för olika nivåer $i = 1, 2, 3$ av den genetiska faktorn (svarande mot att ingen, en eller två muterade varianter ärvts ned från föräldrarna), samt $j = 1, 2, 3, 4$ för graden av motion. För varje nivåkombination i, j deltog $k = 1, \ldots, 3$ personer. Resultatet framgår av följande kvadratsummor, som utgör en del av försökets variansanalystabell:

Variationskälla	Kvs
Gen	4.1
Miljö	13.4
Samspel	18.2
Residual	26.1
Totalt	61.8

a) Utgå från en variansanalysmodell (typ I) med tvåsidig indelning och testa på nivån 5% om det föreligger något samspel mellan arv och miljö. (5 p)

b) Testa på nivån 5% om den genetiska faktorn är signifikant. (5 p)

Uppgift 4

Låt

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

vara en stationär AR(2)-process, med $|\phi_1| + |\phi_2| < 1$, och med oberoende och normalfördelade feltermer $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. Vi antar att $\boldsymbol{X}_T = (X_1, \ldots, X_T)$ är observerad och vill prediktera framtida värden X_{t+k} av processen med $\hat{X}_{T+k} = E(X_{T+k}|\boldsymbol{X}_T)$ för $k = 1, 2, \ldots$. Motsvarande förväntade, kvadratiska prediktionsfel är $\text{MSEP}_k = E[(X_{T+k} - \hat{X}_{T+k})^2]$. (Det antas att ϕ_1 och ϕ_2 är kända, så att det predikterade värdet $\hat{X}_{T+k}$ kan beräknas för $k = 1, 2, \ldots$.)

a) Ge explicita uttryck för $\hat{X}_{T+1}$ and $\hat{X}_{T+2}$. (Ledning: Det enklaste är att först uttrycka X_{T+k} som funktion av $\boldsymbol{X}_T$ och feltermer $\varepsilon_{T+1}, \ldots, \varepsilon_{T+k}$ och därefter bilda $E(X_{T+k}|\boldsymbol{X}_T)$.) (4 p)

b) Använd räkningarna i a) för att beräkna MSEP_1 och MSEP_2 för prediktorerna $\hat{X}_{T+1}$ respektive $\hat{X}_{T+2}$. (4 p)

c) Vad blir gränsvärdet av MSEP_k då $k \to \infty$? (Ledning: Inga långa räkningar krävs, och svaret kan uttryckas med hjälp kovariansfunktionen $\gamma_l = \text{Cov}(X_t, X_{t+l})$, för lämpligt l.) (2 p)

Uppgift 5

Anta att vi har en multipel linjär regressionsmodell

$$Y_i = \tilde{\alpha} + \beta_1(x_{1i} - \bar{x}_1) + \ldots + \beta_m(x_{mi} - \bar{x}_m) + \varepsilon_i, \quad i = 1, \ldots, N \qquad (2)$$

med (centrerat) intercept $\tilde{\alpha}$, effektparametrar $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T$ och oberoende feltermer $\varepsilon_i \sim N(0, \sigma^2)$.

a) Ange ett uttryck för kovariansmatrisen för minsta kvadrat-skattningen $\hat{\boldsymbol{\beta}}$ av $\boldsymbol{\beta}$. (Ledning: Den innehåller σ^2 och $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}$, där $\boldsymbol{X}$ är den del av designmatrisen som härrör från de förklarande variablerna. Både kovariansmatrisen för $\hat{\boldsymbol{\beta}}$, samt matrisen $\boldsymbol{X}$, ska anges.) (3 p)

b) Variansinflationsfaktorn (VIF) är ett mått på hur mycket variansen hos skattningen $\hat{\beta}_j$ av en viss effektparameter β_j ($1 \leq j \leq m$) påverkas av kollineäritet med de andra förklarande variablerna x_k, $k \neq j$. Definiera VIF. (3 p)

c) Visa att i fallet $m = 2$ så ges VIF för den första förklarande variabeln ($j = 1$) i (2) av

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

där R_1^2 är förklaringsgraden i en enkel linjär regression med den första förklarande variabeln $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1N})^T$ i ursprungsmodellen (2) som responsvektor och den andra förklarande variabeln $\boldsymbol{x}_2 = (x_{21}, \ldots, x_{2N})^T$ i ursprungsmodellen (2) som förklarande variabel. (Ledning: Du får utan bevis använda att R_1^2 är kvadraten $\text{Cov}^2(\boldsymbol{x}_1, \boldsymbol{x}_2)/[\text{Var}(\boldsymbol{x}_1)\text{Var}(\boldsymbol{x}_2)]$ på korrelationskoefficienten mellan vektorerna $\boldsymbol{x}_1$ och $\boldsymbol{x}_2$, om dessa tolkas som diskreta fördelningar där varje element i respektive vektor har sannolikhet $1/N$.) (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