

Tentamen för kursen
Linjära statistiska modeller
20 augusti 2025 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 tentamens-tillfället. 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

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

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_i	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 Y_i bland de patienter som hade $x_i = 0$, $x_i = 1$ respektive $x_i = 2$ kopior av den aktuella genvarianten. Syftet med undersökningen var att undersöka om β var positiv.

a) Beräkna minsta kvadrat-skattningen $\hat{\beta}$ av β . (Ledning: Med hjälp av tabellen ovan fås $\sum_{i=1}^{30}(x_i - \bar{x})Y_i = 15$. Vidare kan $\bar{x} = \sum_{i=1}^{30} x_i/30$ och $\sum_{i=1}^{30}(x_i - \bar{x})^2$ beräknas från tabellen.) (3 p)

b) Beräkna $\text{Var}(\hat{\beta})$, 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 i tabellen ovan för att skatta σ^2 . Beräkna därefter medelfelet för skattningen $\hat{\beta}$. (Ledning: Antalet frihetsgrader f för skattningen av σ^2 är detsamma som för variationskällan Inom stickprov för ensidig variansanalys typ I, med tre nivåer (svarande mot antalet kopior av den riskförhöjande genvarianten) på den systematiska faktorn.) (3 p)

d) Beräkna ett ensidigt konfidensintervall för β av typ (a, ∞) med konfidenegrad 97.5%. Ange sedan huruvida vi kan dra slutsatsen att β är positiv. (Ledning: Utnyttja sambandet $t_{0.025}(f) = \sqrt{F_{0.05}(1, f)}$, där kvantilen $F_{0.05}(1, f)$ fås ur tabell och med samma antal frihetsgrader f som i 1c.) (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

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

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 feltermmer 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 för alla tre variationskällorna är $N - 1 = \sum_{i=1}^5 n_i - 1 = 19$. Varje dos i bidrar med $n_i - 1$ frihetsgrader till variationskällan Inom stickprov.) (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 kvadratsumman för linjär regression. Du kommer då ha användning av $\sum_{i,j} (x_{ij} - \bar{x})^2$, där $x_{ij} = i$ är dosen för patient j i grupp i , och där $\bar{x} = \sum_{i,j} x_{ij} / 20$ är den genomsnittliga dosen för alla patienter.) (4 p)

Uppgift 3

En stationär AR(1)-process definieras enligt

$$X_t = \phi X_{t-1} + \varepsilon_t \quad (2)$$

för $t = \dots, -1, 0, 1, \dots$, där $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ är oberoende feltermerna och $|\phi| < 1$.

a) Beräkna kovariansfunktionen $\gamma_k = \text{Cov}(X_t, X_{t+k})$ och korrelationsfunktionen $\rho_k = \text{Corr}(X_t, X_{t+k})$ för $k = 0, 1, 2, \dots$ (Ledning: Använd (2) för att beräkna γ_0 och sedan härleda ett rekursivt uttryck för γ_k för $k = 1, 2, \dots$. Därefter kan ρ_k bestämmas. Du kan utan bevis utnyttja att $\text{Cov}(X_t, \varepsilon_{t+k}) = 0$ för $k = 1, 2, \dots$) (6 p)

b) Anta att T mätningar $\mathbf{X}_T = (X_1, \dots, X_T)$ registrerats av processen. Härled ett uttryck för $\hat{X}_{T+k} = E(X_{T+k} | \mathbf{X}_T)$ för $k = 1, 2, 3, \dots$ (Ledning: Uttryck exempelvis X_{T+k} som en funktion av X_T och feltermerna $\varepsilon_{T+1}, \dots, \varepsilon_{T+k}$. Vid beräkning av $\hat{X}_{T+k}$ kan du även utnyttja att en AR(1)-process är en Markovprocess och alltså ersätta $\mathbf{X}_T$ med X_T i definitionen av $\hat{X}_{T+k}$.) (4 p)

Uppgift 4

Vid en stålindustri försökte man maximera hårdheten hos en viss metall-legering. 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 tillsattes 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

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

för hårdheten hos den legering som erhöles 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 feltermar 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 av att skatta effekten $\Delta = \alpha_2 - \alpha_1 = 2\alpha_2$ av att tillsätta M . Bestäm minsta kvadrat-skattningen $\hat{\Delta} = \hat{\alpha}_2 - \hat{\alpha}_1 = \bar{Y}_{2.} - \bar{Y}_{1.}$ av Δ , om vi vet att $\hat{\Delta} > 0$. (Ledning: $\hat{\Delta}^2$ kan bestämmas utifrån Kvs(Tillsats av M), eftersom varje term $(\bar{Y}_i - \bar{Y}_{..})^2$ i denna kvadratsumma kan uttryckas med hjälp av $\hat{\Delta}^2$.) (3 p)

b) Beräkna en skattning av σ^2 med hjälp av Kvs(Residual). (Ledning: Vid bestämning av antalet frihetsgrader för variationskällan Residual, utnyttja att vi har en tvåsidig variansanalys utan replikat.) (3 p)

c) Beräkna $\text{Var}(\hat{\Delta})$. Bestäm därefter medelfelet hos skattningen $\hat{\Delta}$, samt ange ett 95% konfidensintervall för Δ . (4 p)

Uppgift 5

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

$$Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, \dots, N, \quad (3)$$

där Y_i är responsvariabelns värde för observation i , x_{1i} och x_{2i} 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 feltermar.

a) Skriv modellen på matrisform

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

där $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)$ är en matris med två kolumner $\mathbf{x}_1$ och $\mathbf{x}_2$. Definiera $\mathbf{Y}$, $\mathbf{x}_1$, $\mathbf{x}_2$ och $\boldsymbol{\varepsilon}$. (2 p)

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

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

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

$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}^{-1} = \frac{1}{\alpha\beta - \gamma^2} \begin{pmatrix} \beta & -\gamma \\ -\gamma & \alpha \end{pmatrix}$$

för invertering av den kvadratiske och symmetriska matrisen $\mathbf{X}^T \mathbf{X}$ av ordning 2.) (4 p)

c) Definiera förklaringsgraden R^2 för modellen (3), samt förklaringsgraden R_1^2 och R_2^2 för de två delmodeller av (3) som bara tar med kovariat 1 respektive kovariat 2. Visa sedan att

$$R^2 = R_1^2 + R_2^2 \quad (4)$$

i specialfallet $c = 0$, det vill säga då de två kolumnerna $\mathbf{x}_1$ och $\mathbf{x}_2$ i $\mathbf{X}$ är ortogonala ($\mathbf{x}_1^T \mathbf{x}_2 = 0$). (Ledning: Eftersom modellen (3) inte har med intercept så vill vi inte förklara variationen av Y_i kring $\bar{Y}$, utan kring 0. Ersätt därför $\bar{Y}$ med 0 överallt i den vanliga definitionen av förklaringsgrad.) (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