STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

20 augusti 2025 14–19

Examinator: Ola Hössjer, tel. 070/672 12 18, ola@math.su.se

Uppgift 1

a) Låt $n_j = 10$ beteckna antalet individer med j kopior av den genvariant som antas vara riskförhöjande för diabetes. Totala antalet patienter är $N = n_0 + n_1 + n_2 = 30$. Låt vidare $\bar{Y}_{j\cdot}$ vara medelvärdet av Y_i för de patienter som har j kopior av den aktuella genvarianten. Vi noterar att

$$\begin{array}{rcl} \bar{x} & = & (n_0 \cdot 0 + n_1 \cdot 1 + n_2 \cdot 2)/30 = 10(0+1+2)/30 = 1, \\ \sum_{i=1}^{30}(x_i - \bar{x})^2 & = & 10((-1)^2 + 0^2 + 1^2) = 20, \\ \sum_{i=1}^{30}(x_i - \bar{x})Y_i & = & 10((-1)\bar{Y}_{0\cdot} + 0 \cdot \bar{Y}_{1\cdot} + 1 \cdot \bar{Y}_{2\cdot}) = 10(12.0 - 10.5) = 15. \end{array}$$

Det ger en minsta kvadrat-skattning

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2} = \frac{15}{20} = 0.75.$$

b) Vi har att

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{30}(x_i - \bar{x})^2} = \frac{\sigma^2}{20}.$$

c) Låt s_j vara stickprovsstandardavvikelsen för alla indivder inom gruppen med j kopior av genvarianten. Summan av kvadratavvikelserna $(Y_i - \bar{Y}_{j\cdot})^2$ för individerna i i grupp j, ges av $(10-1)s_j^2 = 9s_j^2$. Den totala kvadratsumman inom alla tre grupper blir

$$\mathrm{Kvs(Inom\ grupp)} = 9(s_1^2 + s_2^2 + s_3^2),$$

och denna variationskälla har $(10-1)+(10-1)+(10-1) = 27$ frihetsgrader. Det ger en skattning

$$\begin{array}{rcl} \hat{\sigma}^2 & = & \mathrm{Mkvs(Inom\ grupp)} \\ & = & \mathrm{Kvs(Inom\ grupp)}/27 \\ & = & (s_1^2 + s_2^2 + s_3^2)/3 \\ & = & (2.0^2 + 2.5^2 + 3.0^2)/3 \\ & = & 6.417 \end{array}$$

av feltermsvariansen. Medelfelet för skattningen av β blir

$$d = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})} = \sqrt{\frac{\hat{\sigma}^2}{20}} = \sqrt{\frac{6.417}{20}} = 0.5664.$$

d) Det följer av a) och c) ovan att ett konfidensintervall av typ (a, ∞) med konfidensgrad 97.5% ges av

$$(\hat{\beta} - t_{0.025}(27)d, \infty) = (0.75 - 2.0518 \cdot 0.5664, \infty) = (-0.412, \infty),$$

där $t_{0.025}(27) = \sqrt{F_{0.05}(1, 27)}$ fås ur tabell. Eftersom 0 ingår i detta intervall kan vi inte på nivån 2.5% förkasta nollhypotesen att den aktuella genvarianten inte har någon riskförhöjande effekt på diabetes. Effekten är för liten för att vara signifikant för ett så pass litet dataset med bara 30 patienter.

Uppgift 2

a) Antalet frihetsgrader för de tre variationskällorna är 1 för Linjär regression (svarande mot skattning av en parameter, β), vidare $5 - 2 = 3$ för Icke-linjäritet (5 parametrar i grundmodellen, en för respektive dos, av vilka 2 skattas i den linära hypotesmodellen), samt slutligen $\sum_{i=1}^{5}(n_i - 1) = N - 5 = 20 - 5 = 15$ för Inom stickprov.

b) Utgående från det beräknade antalet frihetsgrader för Icke-linjäritet och Inom stickprov i a), får vi en

$$\text{F-kvot} = \frac{\text{Mkvs(Icke-linjäritet)}}{\text{Mkvs(Inom stickprov)}} = \frac{\text{Kvs(Icke-linjäritet)}/3}{\text{Kvs(Inom stickprov)}/15} = \frac{35.5/3}{46.0/15} = 3.86,$$

när vi testar den linjära hypotesmodellen mot grundmodellen. Eftersom F-kvoten överstiger tröskelvärdet $F_{0.05}(3, 15) = 3.29$ kan vi förkasta nollhypotesen att det inte finns något icke-linjärt samband mellan dos av medicinen och kreatinhalten, inom det givna intervallet av doser, på signifikansnivån 5%.

c) Under den linjära hypotesmodellen så skattas väntevärdet $\mu_{ij} = E(Y_{ij})$ med

$$\hat{\mu}_{ij} = \hat{\alpha} + \hat{\beta}i = \hat{\alpha} + \hat{\beta}x_{ij}, \quad i = 1, \ldots, 5,\ j = 1, \ldots, n_i,$$

där $x_{ij} = i$. Vidare gäller att

$$\bar{x} = \frac{1}{N}\sum_{i,j} x_{ij} = \frac{1}{N}\sum_{i=1}^{5} n_i \cdot i = \frac{3 \cdot 1 + 3 \cdot 2 + 8 \cdot 3 + 3 \cdot 4 + 3 \cdot 5}{20} = 3,$$

och det centrerade interceptet $\alpha_c = \alpha + \bar{x}\beta = \alpha + 3\beta$ skattas med $\hat{\alpha}_c = \bar{Y}_{\cdot\cdot}$.

Vi får därför att

$$\begin{aligned}\text{Kvs(Linjär regression)} &= \textstyle\sum_{i,j}(\hat{\hat{\mu}}_{ij} - \bar{\bar{Y}}_{\cdot\cdot})^2 \\ &= \textstyle\sum_{i,j}[\hat{\alpha} + \hat{\beta}x_{ij} - (\hat{\alpha} + 3\hat{\beta})]^2 \\ &= \textstyle\sum_{i,j}\hat{\beta}^2(x_{ij} - 3)^2 \\ &= \hat{\beta}^2 \textstyle\sum_i n_i(i-3)^2 \\ &= \hat{\beta}^2[3(-2)^2 + 3(-1)^2 + 8\cdot 0^2 + 3\cdot 1^2 + 3\cdot 2^2] \\ &= 30\hat{\beta}^2.\end{aligned}$$

Eftersom vi vet att $\hat{\beta} < 0$ följer att

$$\hat{\beta} = -\sqrt{\frac{\text{Kvs(Linjär regression)}}{30}} = -\sqrt{\frac{6.8}{30}} = -0.476.$$

Uppgift 3

a) Vi börjar med att bestämma γ_0. Från definitionen av en AR(1)-process följer att

$$\begin{aligned}\gamma_0 &= \text{Var}(X_t) \\ &= \text{Var}(\phi X_{t-1} + \varepsilon_t) \\ &= \phi^2\text{Var}(X_{t-1}) + 2\phi\text{Cov}(X_{t-1}, \varepsilon_t) + \text{Var}(\varepsilon_t) \\ &= \phi^2\gamma_0 + 2\phi\cdot 0 + \sigma_\varepsilon^2 \\ &= \phi^2\gamma_0 + \sigma_\varepsilon^2,\end{aligned} \tag{1}$$

där vi i fjärde steget utnyttjade ledningen. Genom att lösa ut γ_0 ur (1) fås

$$\gamma_0 = \frac{\sigma_\varepsilon^2}{1-\phi^2}. \tag{2}$$

För att bestämma γ_k för $k \geq 1$ används rekursion. Vi har att

$$\begin{aligned}\gamma_k &= \text{Cov}(X_t, X_{t+k}) \\ &= \text{Cov}(X_t, \phi X_{t+k-1} + \varepsilon_{t+k}) \\ &= \phi\text{Cov}(X_t, X_{t+k-1}) + \text{Cov}(X_t, \varepsilon_{t+k}) \\ &= \phi\gamma_{k-1} + 0 \\ &= \phi\gamma_{k-1},\end{aligned} \tag{3}$$

där vi i näst sista steget utnyttjade ledningen. Genom att kombinera (2) med upprepad användning av (3) inses att

$$\gamma_k = \phi^k\phi_0 = \frac{\phi^k\sigma_\varepsilon^2}{1-\phi^2} \tag{4}$$

för $k = 1, 2, \ldots$. För att bestämma autkorrelationsfunktionen så utnyttjas (4). Det ger

$$\rho_k = \text{Corr}(X_t, X_{t+k}) = \frac{\text{Cov}(X_t, X_{t+k})}{\sqrt{\text{Var}(X_t)}\sqrt{\text{Var}(X_{t+k})}} = \frac{\gamma_k}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_k}{\gamma_0} = \phi^k. \tag{5}$$

b) Eftersom en AR(1)-process är en Markovprocess gäller

$$\hat{X}_{T+k} = E(X_{T+k}|\boldsymbol{X}_T) = E(X_{T+k}|X_T). \tag{6}$$

Sedan kan vi antingen utnyttja $E(X_T) = E(X_{T+k}) = 0$, $\text{Var}(X_T) = \text{Var}(X_{T+k}) = \gamma_0$ och det faktum att (X_T, X_{T+k}) är tvådimensionellt normalfördelad, för att i kombination med (6) dra slutsatsen

$$\begin{aligned} \hat{X}_{T+k} &= E(X_{T+k}) + \frac{\sqrt{\text{Var}(X_{T+k})}}{\sqrt{\text{Var}(X_T)}}\text{Corr}(X_T, X_{T+k})(X_T - E(X_T)) \\ &= \text{Corr}(X_T, X_{T+k})X_T = \phi^k X_T. \end{aligned}$$

Alternativt kan vi använda oss av definitionen av en AR(1)-process och skriva om X_{T+k} som

$$X_{T+k} = \phi^k X_T + \phi^{k-1}\varepsilon_{T+1} + \ldots \phi\varepsilon_{T+k-1} + \varepsilon_{T+k}. \tag{7}$$

Insättning av detta uttryck i (6) ger

$$\hat{X}_{T+k} = E\left(\phi^k X_T + \phi^{k-1}\varepsilon_{T+1} + \ldots \phi\varepsilon_{T+k-1} + \varepsilon_{T+k}|X_T\right) = \phi^k X_T, \tag{8}$$

eftersom alla feltermer $\varepsilon_{T+1}, \ldots, \varepsilon_{T+k}$ är oberoende av X_T.

Uppgift 4

a) Minsta kvadratskattningen av $\Delta = \alpha_2 - \alpha_1$ är

$$\hat{\Delta} = \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}.$$

Eftersom $\hat{\Delta} = 2(\bar{Y}_{2\cdot} - \bar{Y}_{\cdot\cdot}) = -2(\bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot})$, så följer att

$$\begin{aligned} \text{Kvs(Tillsats av } M) &= \textstyle\sum_{i=1}^{2}\sum_{j=1}^{6}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= 6[(\bar{Y}_{1\cdot} - \bar{Y}_{\cdot\cdot})^2 + (\bar{Y}_{2\cdot} - \bar{Y}_{\cdot\cdot})^2] \\ &= 6(\hat{\Delta}^2/4 + \hat{\Delta}^2/4)) \\ &= 3\hat{\Delta}^2. \end{aligned}$$

Tillsammans med informatonen $\hat{\Delta} > 0$, så ger det

$$\hat{\Delta} = \sqrt{\frac{\text{Kvs(Tillsats av } M)}{3}} = \sqrt{\frac{3.3}{3}} = 1.049.$$

b) Eftersom vi har en tvåsidig variansanalys utan replikat, och en additiv modell, så har variationskällan Residual $(2-1)(6-1) = 5$ frihetsgrader. Därför gäller att

$$\hat{\sigma}^2 = \text{Mkvs(Residual)} = \frac{\text{Kvs(Residual)}}{5} = \frac{5.2}{5} = 1.04.$$

c) Vi har att

$$\begin{aligned}\mathrm{Var}(\hat{\Delta}) &= \mathrm{Var}(\bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}) \\ &= \mathrm{Var}[\bar{\beta}_{\cdot} + \bar{\varepsilon}_{2\cdot} - (\bar{\beta}_{\cdot} + \bar{\varepsilon}_{1\cdot})] \\ &= \mathrm{Var}(\bar{\varepsilon}_{2\cdot} - \bar{\varepsilon}_{1\cdot}) \\ &= \mathrm{Var}(\bar{\varepsilon}_{1\cdot}) + \mathrm{Var}(\bar{\varepsilon}_{1\cdot}) \\ &= \sigma^2/6 + \sigma^2/6 \\ &= \sigma^2/3.\end{aligned}$$

Det ger ett medelfel

$$d = \sqrt{\widehat{\mathrm{Var}}(\hat{\Delta})} = \sqrt{\frac{\hat{\sigma}^2}{3}} = \sqrt{\frac{1.04}{3}} = 0.589,$$

och ett konfidensintervall

$$\begin{aligned}I_\Delta = &= (\hat{\Delta} - t_{0.025}(5)d, \hat{\Delta} + t_{0.025}(5)d) \\ &= (1.049 - 2.571 \cdot 0.589, 1.049 + 2.571 \cdot 0.589) \\ &= (-0.462, 2.565),\end{aligned} \tag{9}$$

för Δ med konfidensgrad 95%, där $t_{0.025}(5) = \sqrt{F_{0.05}(1,5)}$ kan fås ur tabell. Antalet frihetsgrader 5 för t-fördelningens kvantil svarar mot antalet frihetsgrader för variationskällan Residual i 4b). Eftersom intervallet i (9) innehåller 0 följer att tillsatsen M inte har någon signifikant inverkan på legeringens hållfasthet.

Uppgift 5

a) Modellen skrivs som $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ på matrisform, där $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ är responsvektorn, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ designmatrisen, med första kolumn $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1N})^T$, andra kolumn $\boldsymbol{x}_2 = (x_{21}, \ldots, x_{2N})^T$, samt feltermsvektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T$.

b) Vi inför beteckningarna $s_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j$, $\tilde{\beta}_j$ för minsta kvadrat-skattningen av β_j i en modell där bara kovariat j ingår, samt $\hat{\beta}_j$ för minsta kvadratskattningen av β_j i modellen där båda kovariaterna ingår. I regressionsmodellen $Y_i = \beta_1 x_{1i} + \varepsilon_i$ där bara kovariat 1 ingår så har $\tilde{\beta}_1$ variansen

$$V_1 = \mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^N x_{1i}^2} = \frac{\sigma^2}{s_{11}}.$$

I modellen med båda kovariaterna får vi kovariansmatrisen

$$\mathrm{Var}(\hat{\beta}_1, \hat{\beta}_2) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} = \sigma^2 \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}^{-1}$$

för skattningen av de två effektparametrarna. Eftersom $s_{21} = s_{12}$ så kan vi utnyttja ledningen, och ser att

$$\mathrm{Var}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\sigma^2}{s_{11}s_{22} - s_{12}^2} \begin{pmatrix} s_{22} & -s_{12} \\ -s_{12} & s_{11} \end{pmatrix}. \tag{10}$$

Från första diagonalelementet i denna matris får vi

$$V_2 = \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2 s_{22}}{s_{11}s_{22} - s_{12}^2} = \frac{\sigma^2}{s_{11}(1-c^2)}.$$

Variationsinflationsfaktorn (VIF) anger hur mycket variansen av skattningen av β_1 förstoras på grund av att β_2 måste skattas, dvs

$$\mathrm{VIF} = \frac{V_2}{V_1} = \frac{1}{1-c^2}.$$

c) Med hjälp av ledningen kan vi definiera förklaringsgraden för modellen där bara kovariat 1 ingår, som

$$R_1^2 = \frac{\sum_{i=1}^N \hat{\mu}_i^2}{\sum_{i=1}^N Y_i^2} = \frac{\sum_{i=1}^N (\tilde{\beta}_1 x_{1i})^2}{\sum_{i=1}^N Y_i^2} = \frac{\tilde{\beta}_1^2 s_{11}}{\boldsymbol{Y}^T\boldsymbol{Y}}, \tag{11}$$

eftersom $\mu_i = E(Y_i)$ skattas med $\tilde{\beta}_1 x_{1i}$. Analogt fås att föklaringsgraden för den modell där bara kovariat 2 ingår, är

$$R_2^2 = \frac{\tilde{\beta}_2^2 s_{22}}{\boldsymbol{Y}^T\boldsymbol{Y}}. \tag{12}$$

För modellen med båda kovariaterna har vi att $\hat{\mu}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$. Det ger en förklaringsgrad

$$R^2 = \frac{\sum_{i=1}^N \hat{\mu}_i^2}{\sum_{i=1}^N Y_i^2} = \frac{\sum_{i=1}^N (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})^2}{\boldsymbol{Y}^T\boldsymbol{Y}} = \frac{\hat{\beta}_1^2 s_{11} + \hat{\beta}_2^2 s_{22} + 2\hat{\beta}_1\hat{\beta}_2 s_{12}}{\boldsymbol{Y}^T\boldsymbol{Y}}, \tag{13}$$

där

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} \begin{pmatrix} s_{11}\tilde{\beta}_1 \\ s_{22}\tilde{\beta}_2 \end{pmatrix}.$$

Anta nu att nu att $c = 0$, det vill säga att de två kolumnerna $\boldsymbol{x}_1$ och $\boldsymbol{x}_2$ i $\boldsymbol{X}$ är ortogonala. Av detta följer att $s_{12} = 0$, $\hat{\beta}_1 = \tilde{\beta}_1$ och $\hat{\beta}_2 = \tilde{\beta}_2$. Jämförelse av (11), (12) och (13) ger därför

$$R^2 = R_1^2 + R_2^2.$$