STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

19 augusti 2019 9–14

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

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

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

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

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

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

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) Försöket kan skrivas som en multipel linjär regressionsmodell $\boldsymbol{Y} = \boldsymbol{A\theta} + \boldsymbol{\varepsilon}$, där $\boldsymbol{Y} = (Y_{---}, \ldots, Y_{+++})^T$ innehåller alla responsvärden,

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

är designmatrisen, vars kolumner svarar mot komponenterna i parametervektorn $\boldsymbol{\theta} = (\mu, \bar{K}, \bar{V}, \bar{B}, \overline{KV})^T$, medan $\boldsymbol{\varepsilon} = (\varepsilon_{---}, \ldots, \varepsilon_{+++})^T$ är feltermsvektorn. Minsta kvadrat-skattningen $\boldsymbol{\theta} = (\boldsymbol{AA}^T)^{-1}\boldsymbol{A}^T\boldsymbol{Y} = \boldsymbol{A}^T\boldsymbol{Y}/8$ innehåller skattningar av alla fem parametrar. Vi uttnyttjade här att designmatrisen har ortogonala kolumner, dvs att $\boldsymbol{A}^T\boldsymbol{A} = 8\boldsymbol{I}_5$, där $\boldsymbol{I}_5$ är enhetsmatrisen av ordning 5. För de fyra effektparametrarna får vi MK-skattningar

$$\begin{aligned}\hat{K} &= (-Y_{---} + Y_{+--} - Y_{-+-} + Y_{++-} - Y_{--+} + Y_{+-+} - Y_{-++} + Y_{+++})/8 \\ &= 0.3375,, \\ \hat{V} &= (-Y_{---} - Y_{+--} + Y_{-+-} + Y_{++-} - Y_{--+} - Y_{+-+} + Y_{-++} + Y_{+++})/8 \\ &= 0.5875, \\ \hat{B} &= (-Y_{---} - Y_{+--} - Y_{-+-} - Y_{++-} + Y_{--+} + Y_{+-+} + Y_{-++} + Y_{+++})/8 \\ &= -0.4625, \\ \widehat{KV} &= (Y_{---} - Y_{+--} - Y_{-+-} + Y_{++-} + Y_{--+} - Y_{+-+} - Y_{-++} + Y_{+++})/8 \\ &= -0.0375.\end{aligned} \tag{1}$$

b) Kovariansmatrisen för minsta kvadrat-skattningen $\hat{\boldsymbol{\theta}}$ av $\boldsymbol{\theta}$ är

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \frac{\sigma^2}{8}\boldsymbol{I}_5.$$

Det innebär att alla skattningarna i (1) är oberoende och normalfördelade med samma varians $\sigma^2/8$.

c) Vi skattar σ^2 med hjälp av kvadratsumman för residualerna, vilken kan fås genom att från den totala kvadratsumman subtrahera bort de fyra kvadratsummorna svarande mot var och en av de skattade effektparametrarna. Med hjälp av ledningen och de uträknade skattningarna i a), får vi

$$\begin{aligned}\text{Kvs(Residual)} &= \text{Kvs(Total)} - 8(\hat{K}^2 + \hat{V}^2 + \hat{B}^2 + \widehat{KV}^2) \\ &= 5.529 - 8[0.3375^2 + 0.5875^2 + (-0.4625)^2 + (-0.0375)^2] \\ &= 0.134.\end{aligned}$$

Vi har 8 observationer och 5 regressionsparametrar, dvs $8-5=3$ frihetsgrader för att skatta feltermsvariansen. Det ger en väntevärdesriktig skattning

$$\hat{\sigma}^2 = \text{Mkvs(Residual)} = \frac{\text{Kvs(Residual)}}{3} = \frac{0.134}{3} = 0.0447$$

av σ^2. Från b) ser vi att medelfelet för komponent k i parametervektorn, dvs för θ_k, blir

$$d = \sqrt{\widehat{\text{Var}}(\hat{\theta}_k)} = \sqrt{\frac{\hat{\sigma}^2}{8}} = \sqrt{\frac{0.0447}{8}} = 0.0747.$$

Ett konfidensintervall med konfidensgrad 95% för θ_k ges därför av

$$\hat{\theta}_k \pm t_{0.025}(3)d = \hat{\theta}_k \pm 3.182 \cdot 0.0747 = \hat{\theta}_k \pm 0.238,$$

där $t_{0.025}(3) = \sqrt{F_{0.05}(1,3)}$ kan fås ur tabell. Det betyder att θ_k är signifikant på nivån 5% om $|\hat{\theta}_k| > 0.238$, eftersom dess konfidensintervall då inte täcker över 0. Från a) ser vi att skattningen av de tre huvudeffekterna är signifikanta, däremot inte samspelet mellan kväve och vatten (KV). De tidigare standardnivåerna ger alltså inte optimal tillväxthastighet för växten. Av de åtta försökspunkterna så är det $(+,+,-)$ som ger den högsta skattningen av den förväntade tillväxthastigheten hos växten, dvs om mängden kväve och vatten ökas, medan belysningsmängden sänks.

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 variansnalys 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

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

Det ger ett medelfel

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

och ett konfidensintervall

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

för Δ med konfidensgrad 95%, där $t_{0.025}(5) = \sqrt{F_{0.05}(1,5)}$ kan fås ur tabell. Eftersom detta intervall 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 kvadrat-skattningen 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

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

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

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

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

$$R_1^2 = \frac{\sum_{i=1}^N \hat{\mu}_i^2}{\sum_{i=1}^N Y_i^2} = \frac{\sum_{i=1}^N (\tilde{\beta}_1 x_{1i})^2}{\sum_{i=1}^N Y_i^2} = \frac{\tilde{\beta}_1^2 s_{11}}{\boldsymbol{Y}^T \boldsymbol{Y}}, \tag{3}$$

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

$$R_2^2 = \frac{\tilde{\beta}_2^2 s_{22}}{\boldsymbol{Y}^T \boldsymbol{Y}}. \tag{4}$$

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

$$R^2 = \frac{\sum_{i=1}^N \hat{\mu}_i^2}{\sum_{i=1}^N Y_i^2} = \frac{\sum_{i=1}^N (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})^2}{\boldsymbol{Y}^T \boldsymbol{Y}} = \frac{\hat{\beta}_1^2 s_{11} + \hat{\beta}_2^2 s_{22} + 2\hat{\beta}_1 \hat{\beta}_2 s_{12}}{\boldsymbol{Y}^T \boldsymbol{Y}}. \tag{5}$$

Eftersom

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

så kan täljaren i (5) skrivas om, som

$$\begin{aligned} (\hat{\beta}_1, \hat{\beta}_2) \boldsymbol{X}^T \boldsymbol{X} (\hat{\beta}_1, \hat{\beta}_2)^T &= (s_{11}\tilde{\beta}_1, s_{22}\tilde{\beta}_2)(\boldsymbol{X}^T \boldsymbol{X})^{-1} (s_{11}\tilde{\beta}_1, s_{22}\tilde{\beta}_2)^T \\ &= (s_{11}\tilde{\beta}_1^2 + s_{22}\tilde{\beta}_2^2 - 2s_{12}\tilde{\beta}_1\tilde{\beta}_2)/(1 - c^2) \\ &= (s_{11}\tilde{\beta}_1^2 + s_{22}\tilde{\beta}_2^2 - 2c\sqrt{s_{11}\tilde{\beta}_1^2}\sqrt{s_{22}\tilde{\beta}_2^2})/(1 - c^2), \end{aligned} \tag{6}$$

där vi i andra steget utnyttjade formen för $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ i (2), och i sista steget att $\tilde{\beta}_1$ och $\tilde{\beta}_2$ båda antogs vara positiva. Genom att sätta (6) i (5) ser vi att

$$\begin{aligned} R^2 &= (s_{11}\tilde{\beta}_1^2 + s_{22}\tilde{\beta}_2^2 - 2c\sqrt{s_{11}\tilde{\beta}_1^2}\sqrt{s_{22}\tilde{\beta}_2^2})/[(1-c^2)\boldsymbol{Y}^T\boldsymbol{Y}] \\ &= (R_1^2 + R_2^2 - 2cR_1R_2)/(1-c^2), \end{aligned} \tag{7}$$

där vi i sista steget utnyttjade (3), (4) och att $R_j = \sqrt{R_j^2}$.

d) Vi börjar med att använda (7) för att bestämma korrelationskoefficienten c mellan de två förklarande variablerna. Vi skriver om (7) som en andragradsekvation

$$c^2 - \frac{2R_1R_2}{R_2^2}c + \frac{R_1^2 + R_2^2}{R^2} - 1,$$

som för de givna värdena på R_1^2, R_2^2 och R^2 blir $c^2 - 0.9897c = 0$, med rötter $c_1 = 0$ och $c_2 = 0.9897$. Eftersom vi vet att korrelationskoefficienten är positiv följer att $c = 0.9897$. Från deluppgift b) får vi sedan att VIF $= 1/(1 - 0,9897^2) = 49$.