STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

30 november 2023 8–13

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

Uppgift 1

a) Antalet frihetsgrader för Regression och Residual är 1 respektive 20-2=18. Det följer att

$$\text{F-kvot} = \frac{\text{Mkvs(Regression)}}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Regression)}}{\text{Kvs(Residual)}/18} = \frac{5.2}{19.3/18} = 4.85.$$

Eftersom detta värde överstiger $F_{0.05}(1, 18) = 4.4$, kan vi på signifikansnivån 5% förkasta nollhypotesen att närhet till butik inte har någon inverkan på kundnöjdhet.

b) Skattningen av feltermernas standardavvikelse ges av

$$\hat{\sigma} = \sqrt{\text{Mkvs(Residual)}} = \sqrt{\frac{\text{Kvs(Residual)}}{18}} = \sqrt{\frac{19.3}{18}} = 1.036. \qquad (1)$$

c) Ett 95% konfidensintervall för σ kan skrivas som

$$I_\sigma = \left(\hat{\sigma} \cdot \sqrt{\frac{18}{\chi^2_{0.025}(18)}}, \hat{\sigma} \cdot \sqrt{\frac{18}{\chi^2_{0.975}(18)}} \right) = (0.782, 1.531), \qquad (2)$$

där $\chi^2_p(f)$ är $1-p$-kvantilen för en $\chi^2(f)$-fördelning. Med hjälp av (1)-(2) kan vi bestämma värdena på de två χ^2-kvantiler som ingår i konfidensintervallet för σ. Vi får:

$$\begin{aligned} \chi^2_{0.975}(18) &= 18 \cdot (\hat{\sigma}/1.531)^2 = 8.23, \\ \chi^2_{0.025}(18) &= 18 \cdot (\hat{\sigma}/0.782)^2 = 31.53. \end{aligned}$$

Uppgift 2

a) Parametervektorn för grundmodellen är $\boldsymbol{\theta} = (\tilde{\alpha}_1, \tilde{\alpha}_2, \beta_1, \beta_2, \beta_3)^T$. Hypotesmodellen att det inte finns några systematiska skillnader mellan skadeprissättningen i de två regionerna, svarar mot nollhypotesen $H_0 : \tilde{\alpha}_1 = \tilde{\alpha}_2 = \tilde{\alpha}$. Vi kan formulera det som

$$\boldsymbol{\theta} = \begin{pmatrix} \tilde{\alpha} \\ \tilde{\alpha} \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \tilde{\alpha} \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \boldsymbol{B\lambda}.$$

b) Antalet frihetsgrader för residualerna är $24 - 5 = 19$, och skillnaden i antal parametrar mellan grund- och hypotesmodell är $k - l = 5 - 4 = 1$. För att testa nollhypotesen bildar vi därför

$$\text{F-kvot} = \frac{\text{Mkvs(Avv från hyp)}}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Avv från hyp)}}{\text{Kvs(Residual)}/19} = \frac{1.8}{22.3/19} = 1.53,$$

som är mindre än $F_{0.05}(1, 19) = 4.4$. Vi kan därför inte förkasta nollhypotesen.

c) Grundmodellens underrum spänns upp av de fem kolumnerna i designmatrisen

$$\boldsymbol{A} = (\mathbf{1}_1, \mathbf{1}_2, \boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1, \boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2, \boldsymbol{x}_3 - \bar{\boldsymbol{x}}_3) = (\mathbf{1}_1, \mathbf{1}_2, \boldsymbol{X}),$$

där $\mathbf{1}_1 = (1, \ldots, 1, 0, \ldots, 0)^T$, $\mathbf{1}_2 = (0, \ldots, 0, 1, \ldots, 1)^T$, $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{j,24})^T$, $\bar{\boldsymbol{x}}_j = (\bar{x}_j, \ldots, \bar{x}_j)^T$ och $\boldsymbol{X}$ är den del av designmatrisen som härrör från de förklarande variablerna. Minsta kvadrat-skattningarna av modellparametrarna ges därför av

$$\begin{pmatrix} \hat{\tilde{\alpha}}_1 \\ \hat{\tilde{\alpha}}_2 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = (\boldsymbol{A}^T\boldsymbol{A})^{-1} \begin{pmatrix} \sum_{i=1}^{12} Y_i \\ \sum_{i=13}^{24} Y_i \\ \sum_{i=1}^{24} Y_i(x_{1i} - \bar{x}_1) \\ \sum_{i=1}^{24} Y_i(x_{2i} - \bar{x}_2) \\ \sum_{i=1}^{24} Y_i(x_{3i} - \bar{x}_3) \end{pmatrix}. \qquad (3)$$

Även om $\mathbf{1}_1$ och $\mathbf{1}_2$ är ortogonala mot varandra, är de i allmänhet *inte* ortogonla mot de tre sista kolumnerna $\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1$, $\boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2$, $\boldsymbol{x}_3 - \bar{\boldsymbol{x}}_3$ i designmatrisen. Därför gäller i allmänhet att

$$(\boldsymbol{A}^T\boldsymbol{A})^{-1} \neq \begin{pmatrix} 1/12 & 0 & \mathbf{0}^T \\ 0 & 1/12 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & (\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{pmatrix}, \qquad (4)$$

där $\mathbf{0} = (0, 0, 0)^T$. Av detta och ekvation (3) följer att de två givna formlerna för $\hat{\tilde{\alpha}}_1$ och $\hat{\tilde{\alpha}}_2$ i allmänhet är fel. Observera att *om* $\mathbf{1}_1$ och $\mathbf{1}_2$ hade varit

ortogonala mot kolumnerna i $\boldsymbol{X}$, så hade högerledet i (4) varit lika med $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$, och då hade de föreslagna formlerna för $\hat{\alpha}_1$ och $\hat{\alpha}_2$ varit korrekta.
Hypotesmodellen underrum spänns däremot upp av de fyra kolumnerna i designmatrisen

$$\boldsymbol{A}_0 = \boldsymbol{AB} = (\mathbf{1}, \boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1, \boldsymbol{x}_2 - \bar{\boldsymbol{x}}_2, \boldsymbol{x}_3 - \bar{\boldsymbol{x}}_3),$$

och på grund av definitionen av $\bar{x}_1$, $\bar{x}_2$ och $\bar{x}_3$ i problemtexten är $\mathbf{1} = (1, \ldots, 1)^T$ ortogonal mot de tre sista kolumnerna i $\boldsymbol{A}_0$, det vill säga

$$(\boldsymbol{A}_0^T\boldsymbol{A}_0)^{-1} = \begin{pmatrix} 1/24 & \mathbf{0}^T \\ \mathbf{0} & (\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{pmatrix}.$$

Eftersom $\hat{\hat{\alpha}}$ är första komponenten i minsta kvadrat-skattningen $\hat{\boldsymbol{\lambda}} = (\boldsymbol{A}_0^T\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^T\boldsymbol{Y}$ av regressionsparmetrarna $\boldsymbol{\lambda}$ för hypotesmodellen, följer att

$$\hat{\hat{\alpha}} = \left[(\boldsymbol{A}_0^T\boldsymbol{A}_0)^{-1}\boldsymbol{A}_0^T\boldsymbol{Y}\right]_1 = \frac{1}{24}(\boldsymbol{A}_0^T\boldsymbol{Y})_1 = \frac{1}{24}\sum_{i=1}^{24} Y_i.$$

Uppgift 3

a) Vi börjar med att beräkna

$$\begin{aligned} \gamma_0 &= \mathrm{Var}(X_t) \\ &= \mathrm{Var}(\varepsilon_t - \theta\varepsilon_{t-1}) \\ &= \mathrm{Var}(\varepsilon_t) + \theta^2\mathrm{Var}(\varepsilon_{t-1}) - 2\theta\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1}) \\ &= \sigma_\varepsilon^2 + \theta^2\sigma_\varepsilon^2 - 2\theta \cdot 0 \\ &= (1+\theta^2)\sigma_\varepsilon^2 \end{aligned}$$

och

$$\begin{aligned} \gamma_1 &= \mathrm{Cov}(X_t, X_{t+1}) \\ &= \mathrm{Cov}(\varepsilon_t - \theta\varepsilon_{t-1}, \varepsilon_{t+1} - \theta\varepsilon_t) \\ &= \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+1}) - \theta\mathrm{Var}(\varepsilon_t) - \theta\mathrm{Cov}(\varepsilon_{t-1}, \varepsilon_{t+1}) + \theta^2\mathrm{Cov}(\varepsilon_{t-1}, \varepsilon_t) \\ &= 0 - \theta\sigma_\varepsilon^2 - \theta \cdot 0 + \theta^2 \cdot 0 \\ &= -\theta\sigma_\varepsilon^2. \end{aligned}$$

För $k \geq 2$ gäller att $X_t = \varepsilon_t - \theta_{t-1}\varepsilon_{t-1}$ och $X_{t+k} = \varepsilon_{t+k} - \theta\varepsilon_{t+k-1}$ är funktioner av olika feltermer. Därför är X_t oberoende av X_{t+k} då $k \geq 2$, vilket medför att $\gamma_k = 0$ för $k \geq 2$.

b) Det följer av a) och stationäriteten hos $\{X_t\}$ att

$$\rho_k = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\sqrt{\mathrm{Var}(X_t)}\sqrt{\mathrm{Var}(X_{t+k})}} = \frac{\gamma_k}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_k}{\gamma_0} = \begin{cases} 1, & k = 0, \\ -\theta/(1+\theta^2), & k = 1, \\ 0, & k \geq 2. \end{cases}$$

c) Vi börjar med fallet $k = 1$ och $T = 1$. Eftersom $\boldsymbol{X}_T = X_T$ då $T = 1$ följer att

$$\hat{X}_{T+1} = E(X_{T+1}|\boldsymbol{X}_T) = E(X_{T+1}|X_T) = \rho_1 X_T = -\frac{\theta}{1+\theta^2} X_T,$$

där vi i tredje ledet använde oss av ledningen med $Y = X_T$, $Z = X_{T+1}$, $\mathrm{Var}(Y) = \mathrm{Var}(Z) = \gamma_0 = (1+\theta^2)\sigma_\varepsilon^2$ och $\mathrm{Corr}(Y, Z) = \rho_1$. För $k \geq 2$ och allmänt T kan vi utnyttja att X_{T+k} är oberoende av $\boldsymbol{X}_T$, varav följer att

$$\hat{X}_{T+k} = E(X_{T+k}|\boldsymbol{X}_T) = E(X_{T+k}) = 0.$$

Uppgift 4

a) Antalet observationer är $N = 5$, och antalet parametrar $k = 2$. Designmatrisen

$$\boldsymbol{A} = \begin{pmatrix} 1.3 & 2.6 \\ 0.7 & 2.0 \\ 1.5 & 2.0 \\ 0.9 & 1.4 \\ 2.1 & 2.9 \end{pmatrix}$$

har dimension $N \times k$. Dess två kolumner innehåller värdena på respektive förklarande variabel från alla mätningar. Observera att eftersom modellen saknar intercept har $\boldsymbol{A}$ ingen kolumn med bara ettor.

b) Låt $\boldsymbol{Y} = (Y_1, \ldots, Y_5)^T = (1.6, 1.2, 1.5, 1.2, 1.9)^T$ vara observationsvektorn. Vi använder allmänna formeln för minsta kvadrat-skattningen och ledningen för att räkna ut estimatet

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\hat{\beta}_1, \hat{\beta}_2)^T \\ &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} \\ &= \begin{pmatrix} 9.65 & 15.13 \\ 15.13 & 25.13 \end{pmatrix}^{-1} \begin{pmatrix} 10.24 \\ 16.75 \end{pmatrix} \\ &= \tfrac{1}{9.65 \cdot 25.13 - 15.13^2} \begin{pmatrix} 25.13 & -15.13 \\ -15.13 & 9.65 \end{pmatrix} \begin{pmatrix} 10.24 \\ 16.75 \end{pmatrix} \\ &= \begin{pmatrix} 0.2873 \\ 0.4936 \end{pmatrix} \end{aligned} \tag{5}$$

av $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$.

c) Eftersom antalet frihetsgrader för att skatta feltermsvariansen är $N - k = 5 - 2 = 3$, följer att

$$\hat{\sigma}^2 = \mathrm{Mkvs(Residual)} = \frac{\mathrm{Kvs(Residual)}}{3} = \frac{0.0909}{3} = 0.0303.$$

d) Vi ska utnyttja dualiteten mellan konfidensellipsoider och tester, det vill säga att konfidensellipsoiden för $\boldsymbol{\beta}$ består av mängden av alla $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20})^T$ sådana att nollhypotesen

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$

inte förkastas på signifikansnivån 5%. Under nollhypotesen gäller att $\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\beta}_0$, som vi vill jämföra med $\hat{\boldsymbol{\mu}} = \boldsymbol{A}\hat{\boldsymbol{\beta}}$. Vi ställer därför upp teststorheten

$$\text{F-kvot} = \frac{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2/k}{\hat{\sigma}^2} = \frac{\|\boldsymbol{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2}{2 \cdot \hat{\sigma}^2} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \boldsymbol{C} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \qquad (6)$$

för att testa H_0, med

$$\boldsymbol{C} = \frac{1}{2 \cdot \hat{\sigma}^2} \boldsymbol{A}^T \boldsymbol{A} = \frac{1}{2 \cdot 0.0303} \begin{pmatrix} 9.65 & 15.13 \\ 15.13 & 25.13 \end{pmatrix} = \begin{pmatrix} 159.2 & 249.7 \\ 249.7 & 414.7 \end{pmatrix},$$

och där nollhypotesen förkastas då (6) överstiger $F_{0.05}(k, N-k) = F_{0.05}(2,3) = 9.552$. Genom att kombinera denna ekvation med (5), ser vi att den sökta konfidensellipsoiden ges av

$$\begin{aligned} E \;=\; \{(\beta_{10}, \beta_{20})^T;\; & 159.2(0.287 - \beta_{10})^2 + 2 \cdot 249.7(0.287 - \beta_{10})(0.494 - \beta_{20}) \\ & + 414.7 \cdot (0.494 - \beta_{20})^2 < 9.552\}. \end{aligned}$$

Uppgift 5

a) Det totala stickprovsmedelvärdet är

$$\bar{Y}_{\cdot\cdot} = \frac{1}{5}(\bar{Y}_{1\cdot} + \ldots \bar{Y}_{5\cdot}) = \frac{1}{5}(3.1 + 5.2 + 4.2 + 4.6 + 3.9) = 4.2.$$

Eftersom antal frihetsgrader för variationskällan mellan brädor är $5 - 1 = 4$, ser vi att

$$\begin{aligned} &\text{Mkvs(Mellan brädor)} = \tfrac{1}{4} \textstyle\sum_{i=1}^{5} \sum_{j=1}^{4} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{5} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= (3.1 - 4.2)^2 + (5.2 - 4.2)^2 + (4.2 - 4.2)^2 + (4.6 - 4.2)^2 + (3.9 - 4.2)^2 = 2.46. \end{aligned}$$

b) Vi noterar först att stickprovsvariansen för bräda i ges av

$$s_i^2 = \frac{1}{3} \sum_{j=1}^{4} (Y_{ij} - \bar{Y}_{i\cdot})^2,$$

och att variationskällan inom brädor har $5(4-1) = 15$ frihetsgrader. Det ger

$$\text{Mkvs(Inom brädor)} = \frac{1}{15} \sum_{i=1}^{5} \sum_{j=1}^{4} (Y_{ij} - \bar{Y}_{i\cdot})^2 = \frac{1}{5} \sum_{i=1}^{5} s_i^2 = 0.13.$$

c) Vi ska utnyttja att de två medelkvadratsummorna i a) och b) är observationer av oberoende stokastiska variabler med fördelningar

$$\begin{aligned} \text{Mkvs(Mellan brädor)} &\sim (4\sigma_\delta^2 + \sigma^2)\chi^2(4)/4, \\ \text{Mkvs(Inom brädor)} &\sim \sigma^2 \chi^2(15)/15, \end{aligned}$$

där faktorerna $4\sigma_\delta^2+\sigma^2$ och σ^2 är väntevärdena för respektive kvadratsumma (se formelsamlingen). Av detta följer att

$$\text{F-kvot} = \frac{\text{Mkvs(Mellan brädor)}}{\text{Mkvs(Inom brädor)}} \sim \left(\frac{4}{\sigma^2/\sigma_\delta^2} + 1\right) F(4, 15). \tag{7}$$

Det sökta konfidensintervallet $I_{\sigma^2/\sigma_\delta^2}$ består av de varianskvoter x för vilka ett hypotestest baserat på F-kvoten i (7) inte förkastar nollhypotesen $H_0 : \sigma^2/\sigma_\delta^2 = x$ gentemot den alternativa hypotesen $H_1 : \sigma^2/\sigma_\delta^2 < x$ på signifikansnivån 5%. Ett sådant test förkastar *inte* nollhypotesen då F-kvoten är mindre än $(4/x+1)F_{0.05}(4, 15) = (4/x+1)3.056$. Eftersom det observerade värdet på F-kvoten är $2.46/0.13 = 18.92$ enligt a) och b), följer att

$$I_{\sigma^2/\sigma_\delta^2} = \{x;\, 18.92 < (\frac{4}{x} + 1)3.056\} = (0, \frac{4}{\frac{18.92}{3.056} - 1}) = (0, 0.77).$$