STOCKHOLMS UNIVERSITET,
MATEMATISKA INSTITUTIONEN,
Avd. Matematisk statistik

Lösningsförslag tentamen: Linjära statistiska modeller (MT5001), 2021-12-02

Uppgift 1

(A) Vi får

$$\widehat{\widetilde{\alpha}} = \bar{Y} = 20043.52/200 = 100.2176,$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{44}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{44}(x_i - \bar{x})^2} = \frac{335186.62}{666650.00} = 0.5028.$$

(B) $\widehat{\widetilde{\alpha}}$ är en skattning (MK) baserad på given data för den konstanta men okända parametern $\widetilde{\alpha}$ vilken tolkas som det genomsnittlig företagets förväntade omsättningen.

$\widehat{\beta}$ är en skattning (MK) baserad på given data för den konstanta men okända parametern β vilken tolkas som förväntad ökning i omsättning givet att antalet anställda ökar med en.

(C) Vi har $\bar{x} = 23900.00/200 = 119.5$. Vi får (se Sundberg s 64)

$$\widehat{\sigma}^2 = \frac{1}{200-2}\sum_{i=1}^{200}(Y_i - \widehat{\widetilde{\alpha}} - \widehat{\beta}(x_i - \bar{x}))^2 = 1818.10/198 = 9.1823.$$

Vi får (se Sundberg s 70) konfidensintervallet

$$= \widehat{\widetilde{\alpha}} + \widehat{\beta}(x_0 - \bar{x}) \pm t_{0.05/2}(200-2)\widehat{\sigma}\sqrt{\frac{1}{200} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{200}(x_i - \bar{x})^2}}$$

$$= 100.2176 + 0.5028(x_0 - 119.5) \pm 1.97202\sqrt{9.1823}\sqrt{\frac{1}{200} + \frac{(x_0 - 119.5)^2}{666650}}.$$

(D) Vi får (se Sundberg s 71) simultana konfidensgränser

$$= \widehat{\widetilde{\alpha}} + \widehat{\beta}(x_0 - \bar{x}) \pm \widehat{\sigma}\sqrt{2F_{0.05}(2, 200-2)\left(\frac{1}{200} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{200}(x_i - \bar{x})^2}\right)}$$

$$= 100.2176 + 0.5028(x - 119.5) \pm \sqrt{9.1823}\sqrt{2 \cdot 3.04152\left(\frac{1}{200} + \frac{(x - 119.5)^2}{666650}\right)}.$$

Uppgift 2

(A) Modellen kan t.ex. formuleras som

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

där Y_{ijk} är hastighet för styvhet i, viktfördelning j och slag k. Alltså är α_i effekten av stydhet $i = 1, \ldots, 3$, β_j effekten av viktfördelning $j = 1, \ldots, 3$ och ε_{ijk} feltermerna på slagnivå med $k = 1, \ldots, 10$. Modellen formuleras under bivillkoren $\sum_i \alpha_i = 0$ och $\sum_j \beta_j = 0$ samt under antagande om oberoende och likafördelade $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

(B) Eftersom nivåerna på faktorerna är utvalda inför experimentet, och vi är intresserade av att identifiera eventuella skillnader mellan nivåerna på just dessa faktorer, bör de betraktas som systematiska.

(C) Det krävs att det är mer än en observation per cell (här har vi 10 per cell).

(D) Modellen blir nu t.ex.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

där γ_{ij} är samspelseffekten.

(E) I den övre plotten observerar vi inga besvärliga avvikelser, och drar slutsatsen att antagandet om konstant residualvarians är rimligt. I den undre plotten observerar vi en tydlig negativ trend inom varje cell, eftersom punkterna bildar ett nedåtlutande moln. Detta tyder på en utmattningseffekt inom varje cell (slagen gjordes ju i rad).

Uppgift 3

(A) En multipel linjär regressionsmodell ges av

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ i.i.d.},$$

vilket även kan skrivas som

$$y_i = \tilde{\alpha} + \beta_1 (x_{i1} - \bar{x}_{\cdot 1}) + \beta_2 (x_{i2} - \bar{x}_{\cdot 2}) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ i.i.d.}.$$

(B) Definiera X och S som på s. 79 i Sundberg (2021). Det innebär att

$$\begin{aligned} S &= \begin{pmatrix} \sum_{i=1}^8 (x_{i1} - \bar{x}_{\cdot 1})^2 & \sum_{i=1}^8 (x_{i1} - \bar{x}_{\cdot 1})(x_{i2} - \bar{x}_{\cdot 2}) \\ \sum_{i=1}^8 (x_{i1} - \bar{x}_{\cdot 1})(x_{i2} - \bar{x}_{\cdot 2}) & \sum_{i=1}^8 (x_{i2} - \bar{x}_{\cdot 2})^2 \end{pmatrix} \\ &= \begin{pmatrix} 63.5 & -24.5 \\ -24.5 & 23.5 \end{pmatrix}. \end{aligned}$$

Alltså fås att

$$S^{-1} = \frac{1}{892} \begin{pmatrix} 23.5 & 24.5 \\ 24.5 & 63.5 \end{pmatrix}.$$

Dessutom fås av uppgiftbeskrivningen att

$$\sum_{i=1}^8 (y_i - \bar{y})(x_{i1} - \bar{x}_{\cdot 1}) = -39,$$

$$\sum_{i=1}^8 (y_i - \bar{y})(x_{i2} - \bar{x}_{\cdot 2}) = 17,$$

vilket innebär att

$$X^\top y = \begin{pmatrix} -39 & 17 \end{pmatrix}^\top.$$

Slutligen fås därför att

$$\hat{\beta} = S^{-1}X^\top y = \begin{pmatrix} -0.561 & 0.139 \end{pmatrix}^\top \quad \hat{\hat{\alpha}} = \bar{y} = 5,$$

vilket innebär att

$$\hat{y}_i = 5 - 0.561(x_{i1} - \bar{x}_{\cdot 1}) + 0.139(x_{i2} - \bar{x}_{\cdot 2}).$$

(C) Skattningen ges lämpligen av

$$\frac{1}{5}\sum_{i=1}^{8}\left(y_i - \bar{y} - \sum_{j=1}^{2}\hat{\beta}_j(x_{ij} - \bar{x}_{\cdot j})\right)^2.$$

(D) Konfidensintervallet för komponenterna av $\hat{\beta}$ ges av

$$\hat{\beta}_j + t_{0.025}(5)\hat{\sigma}\sqrt{(S^{-1})_{jj}}.$$

Värden för $(S^{-1})_{jj}$ kan hämtas från deluppgift **(B)** och från tabellen över t-kvantiler fås $t_{0.025}(5) = 2.57$. Dessutom får vi från uppgiftsbeskrivningen att $\hat{\sigma}^2 = 0.355$. Alltså ges konfidensintervallet för β_1 av

$$-0.561 \pm 0.249,$$

och konfidensintervallet för β_2 ges av

$$0.139 \pm 0.409.$$

Konfidensintervallet för β_1 innehåller inte 0 och nollhypotesen kan därför förkastas för denna koefficient. Vad gäller β_2 så kan nollhypotesen inte förkastas eftersom konfidensintervallet innehåller 0.

(E) Antagandena som är lämpliga att granska innefattar att residualerna ska vara approximativt normalfördelade och ha lika varians, att observationerna är oberoende samt att de förklarande variablerna inte ska vara kolinjära. Normalfördelningsantagandet kan exempelvis kontrolleras i en QQ-plot. Att residualerna har lika varians kan förslagsvis verifieras genom att granska hur residualerna varierar beroende på det anpassade värdet. Att observationerna är oberoende kräver att man noga granskar hur experimentet har genomförts och om man har att göra med tidsseriedata kan det vara lämpligt att granska hur residualerna varierar över tid. Avsaknaden av kolinearitet kan t.ex. verifieras genom att beräkna VIF.

Uppgift 4

(A) Med hjälp av lämpliga resultat ifrån formelbladet erhålls enkelt

$$\boldsymbol{c}^T\widehat{\boldsymbol{\theta}} \sim N(\boldsymbol{c}^T\boldsymbol{\theta}, \sigma^2\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{c}).$$

Ifrån ovanstående inses att

$$\frac{\boldsymbol{c}^T\boldsymbol{\theta} - \boldsymbol{c}^T\widehat{\boldsymbol{\theta}}}{\sqrt{\sigma^2\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{c}}} \sim N(0,1).$$

Ifrån formelbladet vet vi även att

$$(N-k)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(N-k).$$

Med hjälp av ovanstående observationer, oberoendet mellan $\widehat{\sigma}^2$ och $\widehat{\boldsymbol{\theta}}$, och definitionen av t-fördelningen (se formlelbladet) inses att

$$\frac{\frac{\boldsymbol{c}^T\boldsymbol{\theta} - \boldsymbol{c}^T\widehat{\boldsymbol{\theta}}}{\sqrt{\sigma^2\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{c}}}}{\sqrt{\frac{(N-k)\widehat{\sigma}^2}{\sigma^2}/(N-k)}} = \frac{\boldsymbol{c}^T\boldsymbol{\theta} - \boldsymbol{c}^T\widehat{\boldsymbol{\theta}}}{\sqrt{\widehat{\sigma}^2\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{c}}} \sim t(N-k)$$

Med grundläggande sannolikhetsteoretiska argument erhålls således det efterfrågade konfidensintervallet.

(B) Enkel linjär regression motsvarar

$$\boldsymbol{A} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_N - \bar{x} \end{pmatrix}$$

och

$$\boldsymbol{\theta} = (\widehat{\alpha}, \beta)^T.$$

Med lite arbete erhålls

$$(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \begin{pmatrix} 1/N & 0 \\ 0 & 1/\sum(x_i - \bar{x})^2 \end{pmatrix}$$

Vi ansätter $\boldsymbol{c} = (1,0)^T$. Vi stoppar in ovanstående i konfidensintervallet i **(A)** och erhåller med lite arbete:

$$\widehat{\widehat{\alpha}} \pm t_{p/2}(N-k)\hat{\sigma}\sqrt{1/N}.$$

Uppgift 5

(A) Se Sundberg s. 259-261.

(B) Se Sundberg s. 271-272. H_0 är att residualerna är sinsemellan oberoende/okorrelerade. Under normalfördelningsantagande gäller approximativt att $\sqrt{T}r_k$ är $N(0,1)$ oberoende. Således gäller approximativt (se s 13) att om vi ansätter r_k^2 och summerar dessa och delar på en lämplig kvantitet så får vi en χ^2-fördelning med ett specifikt antal frihetsgrader (vilket alltså belyser rollen som antagandet om normalfördelning spelar). En variant på detta leder fram till test-statistikan Q och dess fördelning under H_0 (se s 272 för mer information). Regeln för förkastande av H_0 är således att om utfallet Q är större än motsvarande χ^2-kvantil (för vald signifikansnivå) så förkastar vi H_0, vilket i sin tur talar för att modellen eventuellt inte är lämplig för situationen.