STOCKHOLMS UNIVERSITET,
MATEMATISKA INSTITUTIONEN,
Avd. Matematisk statistik

Lösningsförslag tentamen: Linjära statistiska modeller (MT5001), 2021-10-28

Uppgift 1

(A) Vi erhåller

$$\widehat{\widetilde{\alpha}} = \bar{Y} = 399/44 = 9.06818,$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{44}(x_i - \bar{x})Y_i}{\sum_{i=1}^{44}(x_i - \bar{x})^2} = \frac{2602}{28380} = 0.09168.$$

(B) Vi erhåller

$$\widehat{\sigma}^2 = \frac{1}{44-2}\sum_{i=1}^{44}(Y_i - \widehat{\widetilde{\alpha}} - \widehat{\beta}(x_i - \bar{x}))^2 = 50.54/42 = 1.20333.$$

(C) Notera att $\bar{x} = 4092/44 = 93.00$. Vi sätter in siffror i följande formel

$$\widehat{\widetilde{\alpha}} + \widehat{\beta}(100 - \bar{x}) \pm t_{0.05/2}(44-2)\widehat{\sigma}\sqrt{\frac{1}{44} + \frac{(100-\bar{x})^2}{\sum_{i=1}^{44}(x_i - \bar{x})^2}},$$

och erhåller på detta vis konfidensintervallet $[9.36, 10.06]$.

(D) Vi sätter in siffror i följande formel

$$\widehat{\widetilde{\alpha}} + \widehat{\beta}(100 - \bar{x}) \pm t_{0.05/2}(44-2)\widehat{\sigma}\sqrt{1 + \frac{1}{44} + \frac{(100-\bar{x})^2}{\sum_{i=1}^{44}(x_i - \bar{x})^2}},$$

och erhåller på detta vis prediktionsintervallet $[7.47, 11.95]$.

(E) Under antagande att data passar modellen väl så ges väntevärdet av föräljningspriset av en lägenhet om 100 kvm av $\widetilde{\alpha} + \beta(100 - \bar{x}) = \widetilde{\alpha} + 7\beta$. Konstanterna $\widetilde{\alpha}$ och β är dock okända för oss. Konfidensintervallet för detta väntevärdet är det som ges i **(C)**. Prediktionsintervallet i **(D)** ger å andra sidan ett intervall inom vilket försäljningpriset för en lägenhet om 100 kvm kommer att hamna med 95 % sannolikhet. Det senare är större då det beror på osäkerheten som finns i parameterskattningarna (som alltså även finns i konfidensintervallet) men också på osäkerheten som finns i och med att försäljningspriset inte bara beror på lägenhetens storlek utan även på slumpen, dvs pga feltermen.

Uppgift 2

(A) Modellen kan t.ex. formuleras som

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

där Y_{ijk} är betalat belopp för app i, åldergrupp j och betalning k. Vidare är α_i effekten av app $i = 1, \ldots, 5$, β_j effekten av åldersgrupp $j = 1, 2$ och ε_{ijk} feltermerna på betalningsnivå med $k = 1, \ldots, 75 (= 5 \cdot 15)$. 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) Nollhypoteserna är

$$H_0^\alpha : \alpha_i = 0 \, \forall i \quad H_0^\beta : \beta_j = 0 \, \forall j$$

och tolkas som att vi under nollhypotesen inte tror att varken app eller åldersgrupp har en inverkan på betalat belopp. Att vi förkastar H_0^α betyder att vi med statistisk signifikans kan säga att det föreligger skillnad mellan väntevärdena på betalat belopp mellan åtminstone en app och de andra. För H_0^β betyder det att det föreligger en skillnad i väntevärde av betalat belopp mellan unga och äldre (bara 2 nivåer).

(C) Modellen blir nu t.ex.

$$Y_{ijlk} = \mu + \alpha_i + \beta_j + \gamma_{ijl} + \varepsilon_{ijlk}$$

där γ_{ijl} är effekten av person l av åldersgrupp j inom app i. Det tillkomna indexet $l = 1, \ldots, 5$ anger alltså person, och k löper nu från $1, \ldots, 15$. Vi antar att $\gamma_{ijl} \sim N(0, \sigma_\gamma^2)$.

(D) Slutsatsen är att variationen i betalat belopp är större mellan personer än inom personer.

(E) Det är antagandet att $\gamma_{ijl} \sim N(0, \sigma_\gamma^2)$, och mer precist att effekten av person är oberoende av åldersgrupp, som kanske är brutet. Plotten ger vägledning om det då det tillsynes föreligger större variation mellan personer i åldersgrupp "äldre" än i åldersgrupp "unga".

Uppgift 3

(A) Frihetsgraderna i F-testet är $(m, N-m-1)$, där N är antalet observationer och m är antalet förklaringsvariabler. Detta innebär att antalet observationer är lika med $46 + 3 + 1 = 50$.

(B) R^2_{adj} ges av $1 - (1 - R^2)\frac{N-1}{N-m-1} = 1 - (1 - 0.3933)\frac{50-1}{50-3-1} = 0.3537$.

(C) Ett tvåsidigt konfidensintervall för θ_3 ges av $65.88 \pm 2.0129 \times 44.07$, d.v.s. $[-22.8, 154.6]$ (där 2.0129 ges av lämplig t-kvantil, se överst i tentamen). Eftersom 0 finns med i konfidensintervallet kan vi inte förkasta H_0.

(D)

Residualplotten indikerar att feltermernas fördelning är skev åt höger samt att deras varians inte är konstant. Detta innebär att vi inte kan lita på konfidensintervallet (eftersom det beräknades under antagandet om normalfördelade residualer med konstant varians).

(E) Residualplotten indikerar att residualernas varians ökar i takt med det anpassade värdet, vilket är konsistent med en multiplikativ modell. Modellen i uppgiften **(E)** är ekvivalent med den multiplikativa modellen

$$Y_i = \exp\{\theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + x_{i3}\theta_3 + \epsilon_i\} = e^{\theta_0} e^{x_{i1}\theta_1} e^{x_{i2}\theta_2} e^{x_{i3}\theta_3} e^{\epsilon_i}.$$

En log-transformation överför detta multiplikativa samband på en additiv form vilket är önskvärt vid linjär regression.

Uppgift 4

Se Sundbergs kompendium s. 92-93.

Uppgift 5

För definition av begreppet kovarians-stationär se Sundbergs kompendium s. 256. Vi använder att tidsserien är stationär (eftersom $|\theta| < 1$, jmfr. Sundberg s. 258) och erhåller

$$\begin{aligned}
\rho_k &= Corr(Y_t, Y_{t+k}) \\
&= Cov(Y_t, Y_{t+k})/(Std(Y_t)Std(Y_{t+k})) \\
&= Cov(Y_t, Y_{t+k})/V(Y_t).
\end{aligned}$$

Vi använder oberoende för alla ϵ_t och stationäritet för att få

$$\begin{aligned}
Cov(Y_t, Y_{t+1}) &= Cov(Y_t, \phi Y_t + \epsilon_{t+1}) \\
&= Cov(Y_t, \phi Y_t) \\
&= \phi Cov(Y_t, Y_t) \\
&= \phi V(Y_t)
\end{aligned}$$

Vi får ifrån ovan att

$$\rho_1 = \phi.$$

För $k = 2$ noterar vi att

$$\begin{aligned}
Cov(Y_t, Y_{t+2}) &= Cov(Y_t, \phi Y_{t+1} + \epsilon_{t+2}) \\
&= Cov(Y_t, \phi(\phi Y_t + \epsilon_{t+1}) + \epsilon_{t+2}) \\
&= Cov(Y_t, \phi^2 Y_t) \\
&= \phi^2 Cov(Y_t, Y_t) \\
&= \phi^2 V(Y_t)
\end{aligned}$$

vilket ger

$$\rho_2 = \phi^2.$$