STOCKHOLMS UNIVERSITET
MATEMATISK STATISTIK

Lösningar till tentamensskrivning för kursen Linjära statistiska modeller

18 augusti 2023 14–19

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

Uppgift 1

a) Kovariansmatrisen ges av

$$\begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{\sigma^2}{\sum_{i=1}^{20}(x_i-\bar{x})^2} \end{pmatrix}.$$

Eftersom den är diagonal är skattningarna av $\tilde{\alpha}$ och β okorrelerade.

b) Om feltermerna $\varepsilon_i \sim N(0,\sigma^2)$ är normalfördelade så är $(\hat{\tilde{\alpha}},\hat{\beta})^T$ en tvådimensionellt normalfördelad stokastisk variabel med kovariansmatris som i a). Eftersom okorrelerad och oberoende är synonyma begrepp för flerdimensionellt normalfördelade stokastiska variabler så följer då att $\hat{\tilde{\alpha}}$ och $\hat{\beta}$ är oberoende.

c) Med hjälp av ledningen och tabellen med F-kvantiler fås den t-kvantil som behövs för att bilda konfidensintervall för β enligt

$$t_{0.025}(N-2) = t_{0.025}(18) = \sqrt{F_{0.05}(1,18)} = \sqrt{4.4} = 2.10.$$

Feltermernas standardavvikelse skattas med

$$\hat{\sigma} = \sqrt{\text{Mkvs(Residual)}} = \sqrt{0.45} = 0.671.$$

Ett 95% konfidensintervall för β ges därför av

$$I_\beta = \hat{\beta} \pm t_{0.025}(18)\cdot\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{20}(x_i-\bar{x})^2}} = 0.8 \pm 2.10\cdot\frac{0.671}{\sqrt{29.2}} = (0.539, 1.061).$$

Uppgift 2

a) Förklaringsgraden kan uttryckas med hjälp av två kvadratsummor i en variansanalysmodell där grundmodellen testas mot en modell som endast har ett intercept. Det följer att

$$R_0^2 = \frac{\text{Kvs(Regression)}}{\text{Kvs(Total)}} = \frac{\sum_{i=1}^{20}(\hat{\mu}_i-\bar{Y})^2}{\sum_{i=1}^{20}(Y_i-\bar{Y})^2} = \frac{\|\hat{\boldsymbol{\mu}}-\bar{Y}\|^2}{\|\boldsymbol{Y}-\bar{Y}\|^2}, \tag{1}$$

där $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ och $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_N)^T$ är kolumnvektorer med observationer och deras skattade väntevärden enligt grundmodellen, samt $\bar{Y} = (\bar{Y}, \ldots, \bar{Y})^T$.

b) Förklaringsgraden för hypotesmodellen med x_1, x_2 och x_3 som prediktorer fås genom att byta ut $\hat{\mu}_i$ mot $\hat{\hat{\mu}}_i$ i a). Det ger

$$R_1^2 = \frac{\sum_{i=1}^{20}(\hat{\hat{\mu}}_i - \bar{Y})^2}{\sum_{i=1}^{20}(Y_i - \bar{Y})^2} = \frac{\|\hat{\hat{\boldsymbol{\mu}}} - \bar{Y}\|^2}{\|\boldsymbol{Y} - \bar{Y}\|^2}, \tag{2}$$

där $\hat{\hat{\boldsymbol{\mu}}} = (\hat{\hat{\mu}}_1, \ldots, \hat{\hat{\mu}}_N)^T$ är en kolumnvektor med skattade väntevärden för observationerna enligt hypotesmodellen. Genom att ta differensen av (1) och (2) ser vi att

$$R_0^2 - R_1^2 = \frac{\sum_{i=1}^{20}(\hat{\mu}_i - \bar{Y})^2 - \sum_{i=1}^{20}(\hat{\hat{\mu}}_i - \bar{Y})^2}{\sum_{i=1}^{20}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{20}(\hat{\mu}_i - \hat{\hat{\mu}}_i)^2}{\sum_{i=1}^{20}(Y_i - \bar{Y})^2} = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2}{\|\boldsymbol{Y} - \bar{Y}\|^2}, \tag{3}$$

där Pythagoras sats utnyttjades i andra ledet.

c) Ytterligare en användning av Pythagoras sats ger

$$\sum_{i=1}^{20}(Y_i - \bar{Y})^2 = \sum_{i=1}^{20}(Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{20}(\hat{\mu}_i - \bar{Y})^2. \tag{4}$$

Genom att sätta in (4) i uttrycket (1) för grundmodellens förklaringsgrad, ser vi att

$$1 - R_0^2 = \frac{\sum_{i=1}^{20}(Y_i - \bar{Y})^2 - \sum_{i=1}^{20}(\hat{\mu}_i - \bar{Y})^2}{\sum_{i=1}^{20}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{20}(Y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{20}(Y_i - \bar{Y})^2} = \frac{\|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2}{\|\boldsymbol{Y} - \bar{Y}\|^2}. \tag{5}$$

Slutligen kombinerar vi (3) och (5), för att uttrycka

$$\text{F-kvot} = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\hat{\boldsymbol{\mu}}}\|^2/(k-l)}{\|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\|^2/(N-k)} = \frac{(R_0^2 - R_1^2)/(k-l)}{(1-R_0^2)/(N-k)}$$

med hjälp av de två förklaringsgraderna, där $k = 5$ och $l = 4$ är antalet parametrar i grund- respektive hypotesmodellen, medan $N = 20$ anger antalet observationer.

d) Insättning av givna värden från c) ger

$$\text{F-kvot} = \frac{(R_0^2 - R_1^2)}{(1-R_0^2)/(20-5)} = \frac{(0.756 - 0.721)\cdot 15}{1 - 0.756} = 2.15.$$

Eftersom detta värde understiger $F_{0.05}(k-l, N-k) = F_{0.05}(1, 15) = 4.5$ så kan vi inte förkasta hypotesmodellen, och väljer alltså att endast ta med de tre första förklarande variablerna pris, antal konkurrerande företag och hushållens medelinkomst.

Uppgift 3

a) Antalet frihetsgrader för samspel är $(3-1)(4-1) = 6$, och för residualerna $3 \cdot 4(3-1) = 24$. Det ger

$$\text{F-kvot} = \frac{\text{Mkvs(Samspel)}}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Samspel)}/6}{\text{Kvs(Residual)}/24} = \frac{18.2 \cdot 24}{26.1 \cdot 6} = 2.79.$$

Eftersom detta värde överstiger $F_{0.05}(6, 24) = 2.5$ så är samspelet mellan arv och miljö (med knapp marginal) signifikant på nivån 5%.

b) Antalet frihetsgrader för den genetiska faktorn är $3 - 1 = 2$. Det ger

$$\text{F-kvot} = \frac{\text{Mkvs(Gen)}}{\text{Mkvs(Residual)}} = \frac{\text{Kvs(Gen)}/2}{\text{Kvs(Residual)}/24} = \frac{4.1 \cdot 24}{26.1 \cdot 2} = 1.89.$$

Eftersom detta värde understiger $F_{0.05}(2, 24) = 3.4$ så är den genetiska faktorn inte signifikant i sig själv (endast om den kombineras med miljöeffekten enligt a)).

Uppgift 4

a) Eftersom

$$X_{T+1} = \phi_1 X_T + \phi_2 X_{T-1} + \varepsilon_{T+1}$$

följer att

$$\begin{aligned}
\hat{X}_{T+1} &= E[\phi_1 X_T + \phi_2 X_{T-1} + \varepsilon_{T+1} | \boldsymbol{X}_T] \\
&= E[\phi_1 X_T + \phi_2 X_{T-1} + \varepsilon_{T+1} | X_{T-1}, X_T] \\
&= \phi_1 E[X_T | X_{T-1}, X_T] + \phi_2 E[X_{T-1} | X_{T-1}, X_T] + E[\varepsilon_{T+1} | X_{T-1}, X_T] \\
&= \phi_1 X_T + \phi_2 X_{T-1},
\end{aligned}$$

där vi i sista ledet utnyttjade att ε_{T+1} är oberoende av X_{T-1} och X_T (som ju beror av feltermer fram till och med tiden T). För $k = 2$ fås på motsvarande sätt

$$\begin{aligned}
X_{T+2} &= \phi_1 X_{T+1} + \phi_2 X_T + \varepsilon_{T+2} \\
&= \phi_1(\phi_1 X_T + \phi_2 X_{T-1} + \varepsilon_{T+1}) + \phi_2 X_T + \varepsilon_{T+2} \\
&= (\phi_1^2 + \phi_2) X_T + \phi_1 \phi_2 X_{T-1} + \phi_1 \varepsilon_{T+1} + \varepsilon_{T+2}.
\end{aligned}$$

Med liknande räkningar som för $\hat{X}_{T+1}$ ger det

$$\begin{aligned}
\hat{X}_{T+2} &= E[(\phi_1^2 + \phi_2) X_T + \phi_1 \phi_2 X_{T-1} + \phi_1 \varepsilon_{T+1} + \varepsilon_{T+2} | \boldsymbol{X}_T] \\
&= (\phi_1^2 + \phi_2) X_T + \phi_1 \phi_2 X_{T-1}.
\end{aligned}$$

b) Det följer av a) att vi för $k = 1$ och $k = 2$ får prediktionsfelen

$$\begin{aligned}
X_{T+1} - \hat{X}_{T+1} &= \varepsilon_{T+1}, \\
X_{T+2} - \hat{X}_{T+2} &= \phi_1 \varepsilon_{T+1} + \varepsilon_{T+2}.
\end{aligned}$$

Det ger förväntade, kvadratiska prediktionsfel

$$\begin{aligned} \mathrm{MSEP}_1 &= E[\varepsilon_{T+1}^2] = \mathrm{Var}(\varepsilon_{T+1}) = \sigma_\varepsilon^2, \\ \mathrm{MSEP}_2 &= E[(\phi_1\varepsilon_{T+1} + \varepsilon_{T+2})^2] = \mathrm{Var}[\phi_1\varepsilon_{T+1} + \varepsilon_{T+2}] = (\phi_1^2 + 1)\sigma_\varepsilon^2, \end{aligned}$$

där vi i sista ledet utnyttjade att ε_{T+1} och ε_{T+2} är oberoende.

c) Eftersom $\{X_t\}$ är en stationär process gäller att $\rho_k = \mathrm{Corr}(X_t, X_{t+k}) \to 0$ då $k \to \infty$. Då $\{X_t\}$ är en normalprocess så är okorrelerad samma sak som oberoende. Det innebär att X_{T+k} är approximativt oberoende av $\boldsymbol{X}_T$ för stora k, och således gäller approximativt $\hat{X}_{T+k} = E(X_{T+k}|\boldsymbol{X}_T) \approx E(X_{T+k}) = 0$. Det ger ett förväntat kvadratiskt prediktionsfel

$$\mathrm{MSEP}_k = E[(X_{T+k} - \hat{X}_{T+k})^2] \approx E[X_{T+k}^2] = \mathrm{Var}(X_{T+k}) = \gamma_0$$

då k är stor.

Uppgift 5

a) Den del av designmatrisen som härrör från de m förklarande variablerna är

$$\boldsymbol{X} = \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{m1} - \bar{x}_m \\ \vdots & & \vdots \\ x_{1N} - \bar{x}_1 & \dots & x_{mN} - \bar{x}_m \end{pmatrix}.$$

Vidare är

$$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X} = (s_{jk})_{j,k=1}^m$$

en kvadratisk matris av ordning m, med $s_{jk} = \sum_{i=1}^N (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)$. Om kolumnerna i designmatrisen är linjärt oberoende så är $\boldsymbol{S}$ inverterbar. Kovariansmatrisen för effektparameterskattningarna $\hat{\boldsymbol{\beta}}$ ges då av

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{S}^{-1}.$$

b) Variationsinflationsfaktorn för β_j är kvoten mellan variansen $\mathrm{Var}(\hat{\beta}_j)$ för MK-skattningen $\hat{\beta}_j$ och det värde $\mathrm{Var}_0(\hat{\beta}_j) = \sigma^2 s_{jj}^{-1}$ denna varians skulle haft om motsvarande j:te kolumn i $\boldsymbol{X}$ var ortogonal mot dess övriga kolumner. Med andra ord får vi

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{\mathrm{Var}(\hat{\beta}_j)}{\mathrm{Var}_0(\hat{\beta}_j)} = \frac{\sigma^2(\boldsymbol{S}^{-1})_{jj}}{\sigma^2 s_{jj}^{-1}} = s_{jj}(\boldsymbol{S}^{-1})_{jj}.$$

c) Med $m = 2$ förklarande variabler fås

$$\boldsymbol{S}^{-1} = \frac{1}{s_{11}s_{22} - s_{12}^2}\begin{pmatrix} s_{22} & -s_{12} \\ -s_{12} & s_{11} \end{pmatrix}.$$

Det ger

$$\mathrm{VIF}(\hat{\beta}_1) = s_{11}(\boldsymbol{S}^{-1})_{11} = \frac{s_{11}s_{22}}{s_{11}s_{22} - s_{12}^2} = \frac{1}{1 - \frac{s_{12}^2}{s_{11}s_{22}}} = \frac{1}{1 - R_1^2},$$

där vi i sista steget utnyttjade ledningen för att dra slutsatsen

$$R_1^2 = \frac{\mathrm{Cov}^2(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\mathrm{Var}(\boldsymbol{x}_1)\mathrm{Var}(\boldsymbol{x}_2)} = \frac{\left[\frac{1}{N}\sum_{i=1}^N (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]^2}{\frac{1}{N}\sum_{i=1}^N (x_{1i} - \bar{x}_1)^2 \cdot \frac{1}{N}\sum_{i=1}^N (x_{2i} - \bar{x}_2)^2} = \frac{s_{12}^2}{s_{11}s_{22}}.$$