FØRSTE OBLIGATORISKE OPGAVE: MATEMATISK STATISTIK

NIELS RICHARD HANSEN & STEFFEN LAURITZEN

Den første obligatoriske opgave på kurset Matematisk Statistik 2021 består af to komponenter.

- BMS opgave 3.5
- Nedenstående opgave om regression og MLE

Regression og MLE

Lad

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

være tætheden for standard normalfordelingen på $\mathbb{R}$, og lad for $\alpha = (\alpha_0, \alpha_1) \in \mathbb{R}^2$ og fast $z \in \mathbb{R}$

$$f_\alpha(x|z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\alpha_0-\alpha_1 z)^2}$$

betegne tætheden for normalfordelingen med middelværdi $\alpha_0 + \alpha_1 z$ og varians 1. Lad

$$h_\alpha(x, z) = f_\alpha(x|z) f(z). \tag{1}$$

Spørgsmål 1. Vis at h_α er en tæthed for en regulær normalfordeling på $\mathbb{R}^2$ og udtryk middelværdi- og variansparametre i denne fordeling i termer af α_0 og α_1.

Lad nu (X, Z) være todimensionalt normalfordelt med tæthed h_α. Det følger af (1) at

$$f_\alpha(x|z) = \frac{h_\alpha(x, z)}{f(z)}$$

og det er korrekt at tænke på $f_\alpha(\cdot|z)$ som tætheden for den *betingede fordeling* af X givet $Z = z$.

Spørgsmål 2. Lad $(X_1, Z_1), \ldots, (X_N, Z_N)$ være uafhængige og identisk fordelte med samme fordeling som (X, Z). Find log-likelihoodfunktionen og vis at MLE $\hat{\alpha}$ for $\alpha = (\alpha_0, \alpha_1)$ er givet ved (10.28) i EH korollar 10.21 med

$$A = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ \vdots & \vdots \\ 1 & Z_N \end{pmatrix}.$$

Find desuden informationsfunktionen og fisherinformationen og sammenlign (10.29) i EH korollar 10.21 med BMS korollar 4.14.

Vink: Det er vigtigt at forstå, at der overhovedet er noget at vise. Korollar 10.21 giver MLE i en model for X, hvor designmatricen A er fast. I denne opgave bliver I bedt om at finde likelihood og MLE i en model for simultanfordelingen af (X, Z).

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

E-mail address: Niels.R.Hansen@math.ku.dk.

Vi ændrer nu den betingede fordeling til at være

$$f_\beta(x|z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\beta_0-\beta_1 z-\beta_2 z^2)^2},$$

som svarer til en kvadratisk regressionsmodel.

Spørgsmål 3. Vis at $h_\beta(x,z) = f_\beta(x|z)f(z)$ fortsat er en tæthed på $\mathbb{R}^2$. Er det en normalfordeling? Gør også rede for at med $(X_1, Z_1), \ldots, (X_N, Z_N)$ uafhængige og identisk fordelte stokastisk variable med tæthed h_β så er MLE $\hat{\beta}$ for $\beta = (\beta_0, \beta_1, \beta_2)$ ligeledes er givet ved (10.28) i EH korollar 10.21 med

$$A = \begin{pmatrix} 1 & Z_1 & Z_1^2 \\ 1 & Z_2 & Z_2^2 \\ \vdots & \vdots & \vdots \\ 1 & Z_N & Z_N^2 \end{pmatrix}.$$

I det følgende har (X, Z) fordeling med tæthed h_β. Formålet med resten af opgaven er at undersøge konsekvenserne af at fitte en lineær regressionsmodel til observationer fra h.

Spørgsmål 4. Vis at $E(Z) = 0$ og $V(Z) = 1$, og find $E(X)$ og $E(XZ) = \mathrm{Cov}(X, Z)$ udtrykt i termer af β_0, β_1, β_2.

Definer

$$G(\alpha) = E\left((X - \alpha_0 - \alpha_1 Z)^2\right).$$

Spørgsmål 5. Vis at G antager et entydigt minimum, α^{opt}, og udtryk α^{opt} i termer af $E(X)$ og $E(XZ) = \mathrm{Cov}(X, Z)$ såvel som i termer af $\beta = (\beta_0, \beta_1, \beta_2)$.

Vi kan fortolke α^{opt} som de parametre, der giver den bedste affine approksimation af den sande kvadratiske relation, $\beta_0 + \beta_1 Z + \beta_2 Z^2$, mellem X og Z. Med $\hat{\beta}$ MLE som ovenfor fås *plug-in* estimatoren $\hat{\alpha}^{\mathrm{opt}} = \alpha^{\mathrm{opt}}(\hat{\beta})$ for α^{opt}. Fra spørgsmål 2 har vi mindste-kvadraters estimatoren $\hat{\alpha}^{\mathrm{ls}}$, og en tredje naturlig *momentestimator* er

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} X_i \\ X_i Z_i \end{pmatrix}.$$

Spørgsmål 6. Vis at såvel $\hat{\alpha}$ som $\hat{\alpha}^{\mathrm{opt}}$ er asymptotisk normalfordelte med asymptotisk middelværdi α^{opt} og find de asymptotiske variansmatricer. Sammenlign med variansmatricen for mindste-kvadraters estimatoren $\hat{\alpha}^{\mathrm{ls}}$ fra spørgsmål 2, og overvej hvorvidt den variansmatrix er korrekt, når data nu er fra h_β.
Vink: Den asymptotiske varians for $\hat{\alpha}$ kan udtrykkes i termer af momenter. Det er muligt at udtrykke disse i termer af β, men det er ikke strengt nødvendigt at lave de udregninger. I praksis, f.eks. i simulationerne nedenfor, kan I bruge empiriske momenter til beregning af standard error.

Spørgsmål 7. Udfør et simulationsstudie for at undersøge fordelingen af $\hat{\alpha}_1$, $\hat{\alpha}_1^{\mathrm{opt}}$ og $\hat{\alpha}_1^{\mathrm{ls}}$. Undersøg i særdeleshed hvordan forskellige teoretiske resultater om standard error for disse estimatorer stemmer overens med simulationerne. Prøv forskellige valg af $\beta_0, \beta_1, \beta_2$ og N.