Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1, vejledende besvarelse

14. april 2016

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `april16.R`.

Opgave 1

1. Likelihoodfunktionen og log-likelihoodfunktionen (på nær en additiv konstant) for en observation x er givet ved

$$\begin{aligned} L_x(\beta,\sigma^2) &= \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}\big(x_i - e^{\beta t_i}\big)^2\Big) \\ &\propto \big(\sigma^2\big)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum\big(x_i - e^{\beta t_i}\big)^2\Big) \\ \ell_x(\beta,\sigma^2) &= -\log L_x(\beta,\sigma^2) = \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\sum\big(x_i - e^{\beta t_i}\big)^2 \end{aligned}$$

   hvor $(\beta,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+$, og produktet samt summerne er fra $i = 1$ til n.

   Scorefunktionen fås ved differentiation mht. β hhv. σ^2:

$$S_x(\beta,\sigma^2) = \left(-\frac{1}{\sigma^2}\sum\big(x_i - e^{\beta t_i}\big)e^{\beta t_i}t_i \quad , \quad \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\sum\big(x_i - e^{\beta t_i}\big)^2\right)$$

   Indsættelse af de foreslåede værdier giver følgende:

```
> n <- 6
> s2 <- 1.606481
> b <- 1.15625068
> sum((x-exp(b*t))*exp(b*t)*t)              # Første koordinat i S(b,s2)
[1] 6.034892e-05
> n/2/s2 - sum((x-exp(b*t))^2) / 2 / s2^2   # Anden koordinat i S(b,s2)
[1] 3.051034e-07
```

   Altså er $S_x(1.15625068, 1.606481)$ lig nul pånær afrunding.

2. Elementet på plads (1,2) i den stokastiske version af informationsfunktionen er

$$I_{X,12}(\beta,\sigma^2) = \frac{1}{\sigma^4}\sum(X_i - e^{\beta t_i})e^{\beta t_i}t_i$$

   der har middelværdi nul eftersom $\mathrm{E}X_i = e^{\beta t_i}$. Således er element (1,2) i Fisherinformationen lig nul.

   Elementet på plads (1,1) i den stokastiske version af informationsfunktionen er

$$I_{X,11} = -\frac{1}{\sigma^2}\sum\Big\{-e^{2\beta t_i}t_i^2 + (X_i - e^{\beta t_i})e^{\beta t_i}t_i^2\Big\}$$

   Her har sidste led middelværdi nul, så vi får

$$i(\beta,\sigma^2)_{11} = \mathrm{E}I_{X,11} = \frac{1}{\sigma^2}\sum t_i^2 e^{2\beta t_i}$$

3. Det sædvanlige asymptotiske resultat giver at $\hat{\beta}$ er asymptotisk normalfordelt med middelværdi β og varians lig element $(1,1)$ i den inverse Fisherinformation.

   Eftersom Fisherinformationen er en diagonalmatrix, er den asymptotiske varians altså

   $$\left[i(\beta,\sigma^2)^{-1}\right]_{11} = \frac{1}{i(\beta,\sigma^2)_{11}} = \frac{\sigma^2}{\sum t_i^2 e^{2\beta t_i}}$$

   Ved at indsætte estimaterne, fås den estimerede spredning

   $$\mathrm{SE}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum t_i^2 e^{2\hat{\beta} t_i}} = 0.0121,$$

   og det falske 95% Wald konfidensinterval er så

   $$\hat{\beta} \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}) = 1.156 \pm 0.024 = \big(1.132\,,\, 1.180\big).$$

4. Den alternative model er en lineær normal model for Z'erne. Designmatricen A er en $1 \times n$ matrix med t_i på plads i. Vi får derfor

   $$\hat{\gamma} = (A^T A)^{-1} A^T Z = \frac{\sum t_i z_i}{\sum t_i^2}, \quad \mathrm{Var}(\hat{\gamma}) = \tau^2 (A^T A)^{-1} = \frac{\tau^2}{\sum t_i^2}, \quad \mathrm{SE}(\hat{\gamma}) = \frac{\tilde{\tau}}{\sqrt{\sum t_i^2}}.$$

   For det givne datasæt får vi

   $$\hat{\gamma} = 1.1454, \quad \tilde{\tau}^2 = 0.2874, \quad \mathrm{SE}(\hat{\gamma}) = 0.1205.$$

   Bemærk evt. at den alternative model er en lineær regression med respons x og forklarende variabel t, men uden intercept. Modellen kan fittes med kommandoen `lm(z~t-1)`. Fra summary for denne model kan estimat og estimeret spredning aflæses direkte.

Opgave 2

1. Vi har $Y = CX$ hvor

   $$C = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 2 & -3 \end{pmatrix}$$

   Det følger at

   $$Y \sim N(0, C\mathrm{V}X C^T),$$

   og når man regner på variansmatricen får man

   $$\mathrm{V}Y = \begin{pmatrix} 10 & -24 \\ -24 & 63 \end{pmatrix}$$

   Denne matrix har determinant 54 og er dermed invertibel, så fordelingen af Y er en regulær normalfordeling på $\mathbb{R}^2$.

2. Betragt så

   $$\begin{pmatrix} Z \\ X_1 + X_3 \end{pmatrix} = \begin{pmatrix} c_1 & c_3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$$

   der er normalfordelt med på $\mathbb{R}^2$ med middelværdi 0 og variansmatrix

   $$\Gamma = \begin{pmatrix} c_1 & c_3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 & 1 \\ c_3 & 1 \end{pmatrix} = \begin{pmatrix} c_1^2 + 4c_3^2 + 2c_1c_3 & 2c_1 + 5c_3 \\ 2c_1 + 5c_3 & 7 \end{pmatrix}$$

Kravene om at $VZ = 21$ og at Z og $X_1 + X_3$ er uafhængige er derfor opfyldt hvis og kun hvis

$$c_1^2 + 4c_3^2 + 2c_1c_3 = 21, \quad 2c_1 + 5c_3 = 0.$$

Fra den sidste betingelse får $c_1 = -\frac{5}{2}c_3$, der indsættes i første betingelse:

$$\frac{25}{4}c_3^2 + 4c_3^2 - 5c_3^2 = 21 \Leftrightarrow \frac{21}{4}c_3^2 = 21 \Leftrightarrow c_3^2 = 4 \Leftrightarrow c_3 = \pm 2$$

For $c_3 = -2$ fås $c_1 = 5$, for $c_3 = 2$ fås $c_1 = -5$, så betingelserne er opfyldt for $(c_1, c_3) = (5, -2)$ og $(c_1, c_3) = (-5, 2)$.

Opgave 3

1. Lad $X_1, \ldots, X_{50}$ være de stokastiske variable svarende til log-serinmængden. I `modelA` antages $X_!, \ldots, X_n$ at være uafhængige, og $X_i \sim N(\xi_i, \sigma^2)$ hvor

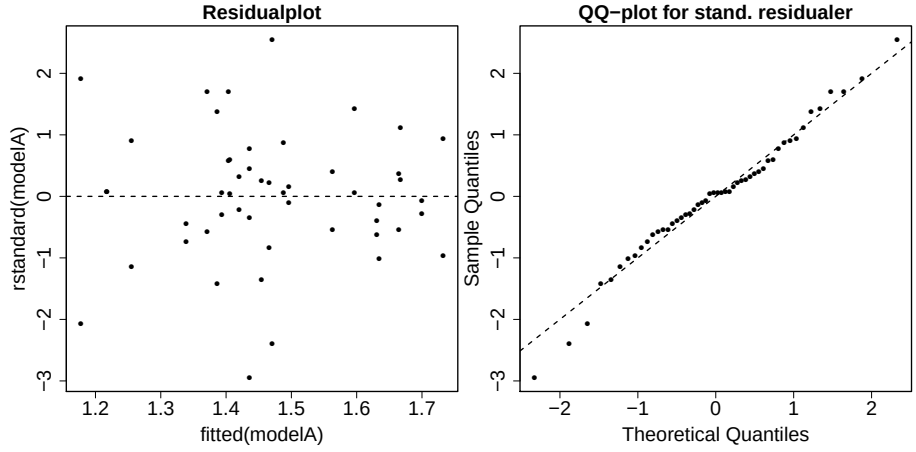
   $$\xi_i = \alpha_{f(i)} + \beta_{f(i)} \cdot t_i$$

   hvor t_i er hydrolysetiden, $f(i)$ angiver fodertypen for observation i, α'erne er ukendte interceptparametre og β'erne er ukendte hældningsparametre. Der er fem α'er og fem β'er svarende til at der både intercept og hælding er forskellige for de fem fodertyper.

   Residualplot og normalfordelings-QQ-plot for de standardiserede residualer er vist nedenfor. Vi noterer følgende:

   - Punkterne i residualplottet fordeler sig ligeligt om x-aksen og med cirka lige stor lodret variation i alle dele af plottet (fra venstre mod højre). Der er således ingen tegn på misspecifikation af middelværdi eller på variansinhomogenitet.
   - Punkterne i QQ-plottet varierer omkring den rette linie ($y = x$), så normalfordelingsantagelsen synes at være fornuftig. Der er heller ikke kritiske outliers.

   Vi har desuden ikke grund til at tvivle på antagelsen om uafhængighed. Samlet ser ser `modelA` altså ud til at være en fornuftig model til data.



2. Hypotesen er

   $$H : \beta_{\texttt{byg}} = \beta_{\texttt{fiskemel}} = \beta_{\texttt{majs}} = \beta_{\texttt{kb.mel}} = \beta_{\texttt{soja}}$$

Den statistiske model under hypotesen er præcis `modelB`.

Hypotesen testes derfor ved at sammenligne modellerne `modelA` og `modelB`. Testet udføres som et F-test. F-teststørrelsen er 2.07 der skal vurderes i F-fordelingen med $(4,40)$ frihedsgrader. Dette giver p-værdien 0.10, så vi kan ikke forkaste hypotesen.

Data tyder således ikke på at hældningerne er forskellige, og det er rimeligt at fortsætte analysen vha. `modelB` hvor hældningerne er ens.

3. Vi får følgende estimater og konfidensintervaller:

   - Forskellen mellem forventet log-serinmængde for byg og majs, dvs. $\delta_{\texttt{bm}} = \alpha_{\texttt{majs}} - \alpha_{\texttt{byg}}$. Estimat og konfidensinterval kommer direkte fra paramteriseringen i `modelB`:

   $$\hat{\delta}_{\texttt{bm}} = 0.158, \quad 95\%\text{ KI } (0.149, 0.167)$$

   - Forskellen mellem forventet log-serinmængde for majs og soja, dvs. $\delta_{\texttt{ms}} = \alpha_{\texttt{soja}} - \alpha_{\texttt{majs}}$. Estimat og konfidensinterval kan fx fås ved at ændre referencegruppe. Vi får

   $$\hat{\delta}_{\texttt{ms}} = 0.071, \quad 95\%\text{ KI } (0.062, 0.080)$$

4. Forskellen mellem forventet log-serinmængde ved 20 og 50 timers hydrolysetid (for fastholdt fodertype), dvs. 30β hvor β er den fælles hældning. Estimat og KI fås ved at gange estimat og KI for β med 30:

   $$30\hat{\beta} = -0.122, \quad 95\%\text{ KI } (-0.126, -0.118)$$

   Den prædikterede værdi for log-serin efter 40 timer for soja er

   $$1.535 + 0.229 - 0.00407 \cdot 40 = 1.602$$

   hvor estimaterne er taget fra `modelB`. 95% Prædiktionsintervallet fås nemmest vha. `predict`-funktionen og viser sig at være $(1.581\,,\, 1.623)$.

5. For fodertype j er $\gamma_j = \alpha_j + 50\beta$. Således er

   $$\bar{\gamma} = \frac{1}{5}\sum_{j=1}^{5} \gamma_j = \frac{1}{5}\sum_{j=1}^{5} \alpha_j + 50 \cdot \beta.$$

   Hvis vi sætter $\theta = (\alpha_1, \ldots, \alpha_5, \beta)$ og $\psi^T = (1/5, 1/5, 1/5, 1/5, 1/5, 50)$, har vi altså

   $$\bar{\gamma} = \psi^T \theta.$$

   Vi benytter eksempel 10.30, og får

   $$\hat{\bar{\gamma}} = \psi^T \hat{\theta}, \quad \mathrm{Var}(\hat{\bar{\gamma}}) = \psi^T \mathrm{Var}(\hat{\theta}) \psi$$

   Den ønskede parametrisering opnås med kommandoen

   ```
   lm(log(serin) ~ foder+tid-1, data=hydrolyse)
   ```

   og estimat $\hat{\theta}$ og variansmatrix $\mathrm{Var}(\hat{\theta})$ fås fx. med `coef` og `vcov`. Vi får $\hat{\bar{\gamma}} = 1.39236$ med estimeret spredning $\mathrm{SE}(\hat{\bar{\gamma}}) = 0.00185$.

   Det tilhørende 95% konfidensinterval er

   $$1.39236 \pm 2.015 \cdot 0.00185 = 1.39236 \pm 0.00373 = (1.38863, 1.39609)$$

   hvor vi har benyttet at 97.5% fraktilen i t_{44} fordelingen er 2.015.

Opgave 4

1. Scorefunktionen og informationsfunktionen er

$$S_x(\theta) = -\frac{n}{\theta} + \frac{\sum x_i}{1-\theta}, \quad I_x(\theta) = \frac{n}{\theta^2} + \frac{\sum x_i}{(1-\theta)^2}.$$

   Vi løser scoreligningen

$$S_x(\theta) = 0 \Leftrightarrow \theta \sum x_i = n - n\theta \Leftrightarrow \theta(n + \sum x_i) = n \Leftrightarrow \theta = \frac{n}{n + \sum x_i},$$

   og bemærker desuden at $I_x(\theta) > 0$ for alle $\theta \in (0,1)$. Løsningen ligger i $(0,1]$ og er kun 1 hvis $\sum x_i = 0$ (hvilket sker med θ^n). Hvis vi tillader en estimator på randen, er løsningen til scoreligningen således en entydigt bestemt MLE, dvs.

$$\hat{\theta} = \frac{n}{n + \sum x_i}.$$

2. For den givne observation er $\sum x_i = 33$, så ML estimatet er

$$\hat{\theta} = \frac{12}{12+33} = 0.267.$$

   Likelihood ratio teststørrelsen er

$$LR(x) = 2\ell_x(0.4) - 2\ell_x(\hat{\theta}) = 3.51$$

   der skal vurderes i χ^2 fordelingen med en frihedsgrad. Dette giver p-værdien 0.06. Med et signifikansniveau på 5% kan vi således ikke afvise hypotesen.

3. Det udfyldte skema ser således ud (varierer naturligvis lidt fra simulation til simulation):

n	θ	Relativ hyppighed
20	0.4	0.052
20	0.5	0.255
20	0.6	0.713
40	0.4	0.049
40	0.5	0.458
40	0.6	0.953

   Vi ser at

   - Det faktiske niveau for testet er tæt på de ønskede 5%, både for $n = 20$ og $n = 40$
   - Styrken vokser når n vokser (som forventet)
   - Styrken vokser når afstanden mellem det sande θ og 0.4 vokser (som forventet)