Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1, vejledende besvarelse

29. juni 2017

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `august17.R`.

Opgave 1

1. Likelihoodfunktionen er (proportional med)

$$L_x(\theta) = \prod_{i=1}^n \theta t_i x_i^{\theta t_i - 1} \propto \theta^n \prod_{i=1}^n x_i^{\theta t_i}$$

Vi får således (på nær en additiv) konstant

$$\ell_x(\theta) = -\log L_x(\theta) = -n\log\theta - \theta\sum_{i=1}^n t_i \log x_i$$

og dermed

$$S_x(\theta) = \ell_x'(\theta) = -\frac{n}{\theta} - \sum_{i=1}^n t_i \log x_i$$
$$I_x(\theta) = S_x'(\theta) = \frac{n}{\theta^2}$$
$$i(\theta) = \mathrm{E}_\theta I_X(\theta) = \frac{n}{\theta^2}$$

2. Vi løser først scoreligningen for en observation x:

$$S_{x,y}(\theta) = 0 \Leftrightarrow \frac{n}{\theta} = -\sum_{i=1}^n t_i \log x_i \Leftrightarrow \theta = -\frac{n}{\sum_{i=1}^n t_i \log x_i}$$

Der er således et entydigt stationært punkt. Da der desuden gælder $I_x(\theta) > 0$ for alle $\theta > 0$, giver det stationære punkt anledning til et minimum for $\ell_{x,y}$. Bemærk at løsningen til scoreligningen er positiv da alle $x_i \in (0,1)$, så løsningen ligger i parametermængden.

Ovenstående gælder for alle $x \in (0,1)^n$, så vi får at ML estimatoren er

$$\hat\theta = -\frac{n}{\sum_{i=1}^n t_i \log X_i}.$$

Den asympotiske fordeling af $\hat\theta$ er

$$\hat\theta \overset{as}{\sim} N\big(\theta, i(\theta)^{-1}\big), \text{ dvs. } \hat\theta \overset{as}{\sim} N\Big(\theta, \frac{\theta^2}{n}\Big)$$

3. Lad $t > 0$ være givet og lad X have tæthed $f(x) = \theta t \cdot x^{\theta t - 1}$ for $x > 0$. Definer desuden funktionen $h : (0,1) \to (0,\infty)$ ved $h(x) = -t \log x$ og $Y = h(X)$. Funktionen h er strengt aftagende, og dermed bijektiv, samt kontinuert differentiabel. Vi får følgende:

$$h^{-1}(y) = e^{-y/t}, \quad Dh^{-1}(y) = -\frac{1}{t}e^{-y/t}, \quad y > 0,$$

og det følger fra den endimensionale transformationssætning af tætheden for Y er givet ved $g(y) = 0$ for ≤ 0 og

$$g(y) = f\big(h^{-1}(y)\big)\big|Dh^{-1}(y)\big| = \theta t \big(e^{-y/t}\big)^{\theta y - 1}\frac{1}{t}e^{-y/t} = \theta e^{-\theta y}$$

for $y > 0$. Således er Y eksponentialfordelt med middelværdi $1/\theta$ eller gammafordelt med formparameter 1 og skalaparameter $1/\theta$: $Y_i \sim \Gamma(1, 1/\theta)$.

Altså er $Y_1, \ldots, Y_n$ uafhængige (fordi X_i'erne er det) og alle $Y_i \sim \Gamma(1, 1/\theta)$. Pga. foldningsegenskaben for gammafordelingen, får vi så $S_Y \sim \Gamma(n, 1/\theta)$ og endelig $\theta S_Y \sim \Gamma(n, 1)$.

Således er θS_Y en pivot, og hvis g_1 og g_2 er 2.5% og 97.5% fraktilerne i $\Gamma(n, 1)$ fordelingen, så er

$$0.95 = P(g_1 < \theta S_Y < g_2) = P\Big(\frac{g_1}{S_Y} < \theta < \frac{g_2}{S_Y}\Big)$$

og

$$\Big(\frac{g_1}{S_Y}, \frac{g_2}{S_Y}\Big)$$

er et eksakt 95% konfidensinterval for θ.

4. Vi har fra tidligere at

$$\log L_x(\theta) = n\log\theta + \theta\sum_{i=1}^{n} t_i \log x_i = n\log\theta - n\theta\bar{y},$$

og at $\hat{\theta} = 1/\bar{y}$. Derfor er

$$\begin{aligned} LR(\theta, x) &= 2\Big(\log L_x(\hat{\theta}) - \log L_x(\theta)\Big) \\ &= 2\Big(-n\log\bar{y} - n - n\log\theta + n\theta\bar{y}\Big) \\ &= 2n\Big(-\log\bar{y} - 1 - \log\theta + \theta\bar{y}\Big). \end{aligned}$$

For hypotesen $H : \theta = 4$ fås $LR(4, x) = 1.83$. Hvis vi benytter χ^2_1 approksimationen til fordelingen af $LR(\theta, X)$ under hypotesen, fås p-værdien $P\big(LR(X, 4) \geq 1.83\big) = 0.18$, så vi kan ikke afvise hypotesen: Der er altså ikke evidens i data for at θ er forskellig fra 4.

5. For det givne datasæt og $n = 10$ er

$$S_y = 1.578, \quad \bar{y} = 0.1578, \quad \hat{\theta} = 6.337, \quad g_1 = 4.795, \quad g_2 = 17.08$$

Det eksakte 95% konfidensinterval er således

$$\Big(\frac{g_1}{S_y}, \frac{g_2}{S_y}\Big) = (3.039, 10.827).$$

Konfidensintervallet baseret på $LR(\theta, X)$ er defineret ved

$$\{\theta > 0 \,|\, LX(\theta, x) < q_{0.95}\}$$

hvor $q_{0.95} = 3.84$ er 95% fraktilen i χ^2 fordelingen med en frihedsgrad. Eftersom $\theta \to LR(\theta, x)$ er konveks og $LR(\hat{\theta}, x) = 0$, er endepunkterne i konfidensintervallet løsningerne til ligningen $LR(\theta, x) = 3.841$. Hvis vi indsætter $\theta = 3.1754$ og $\theta = 11.1151$ i udtrykket for $LR(\theta, x)$ får vi netop 3.841 (på nær afrundingsfejl).

6. Den asymptotiske fordeling af $\hat{\theta}$ er $N(\theta, \theta^2/n)$, specielt er spredning for $\hat{\theta}$ approximativt $\theta/\sqrt{n}$. Jeg fik følgende skema med 5000 simulationer:

		Simulation		Asymptotisk fordeling	
n	θ	middelværdi	spredning	gennemsnit	spredning
10	5	5.59	1.99	5	1.58
25	5	5.21	1.08	5	1
250	5	5.02	0.32	5	0.32

Vi ser at middelværdi og spredning i den asympotiske fordeling først er fornuftige for $n = 250$. For $n = 10$ og $n = 25$ er den faktiske middelværdi og den faktiske varians begge større end den asymptotiske fordeling tilsiger.

Specielt er $\hat{\theta}$ ikke en central estimator, thi så skulle den have den korrekte middelværdi også for små værdier af n. (Derimod er $1/\hat{\theta}$ faktisk central for $1/\theta$, hvoraf det i øvrigt via Jensens ulighed følger at $\hat{\theta}$ ikke er central for θ.)

Opgave 2

1. Hvis vi definerer

$$C = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix},$$

så er $Y = CX$, og tranformationssætningen for normalfordelingen giver

$$Y = CX \sim N(0, C\Sigma C^T).$$

Variansmatricen viser sig at være

$$VY = C\Sigma C^T = \begin{pmatrix} 13 & 19 \\ 19 & 30 \end{pmatrix}.$$

Variansmatricen har determinant 29 og er dermed regulær, så Y er regulært normalfordelt på $\mathbb{R}^2$.

2. Vi har at $\det(\Sigma) = 0$, så X er singulært normalfordelt på $\mathbb{R}^3$.

Sæt $D = \begin{pmatrix} 1 & -1 & -1 \end{pmatrix}$. Så er $Z = X_1 - X_2 - X_3 = DX \sim N(0, D\Sigma S^T) = N(0,0)$. Altså er $Z = 0$, eller $X_3 = X_1 - X_2$, med sandsynlighed 1, dvs. $P(X \in V) = 1$. Det er desuden klart at X ikke kan være koncentreret på en mængde af lavere dimension eftersom fx $(X_1, X_2)^T$ er regulært normalfordelt på $\mathbb{R}^2$.

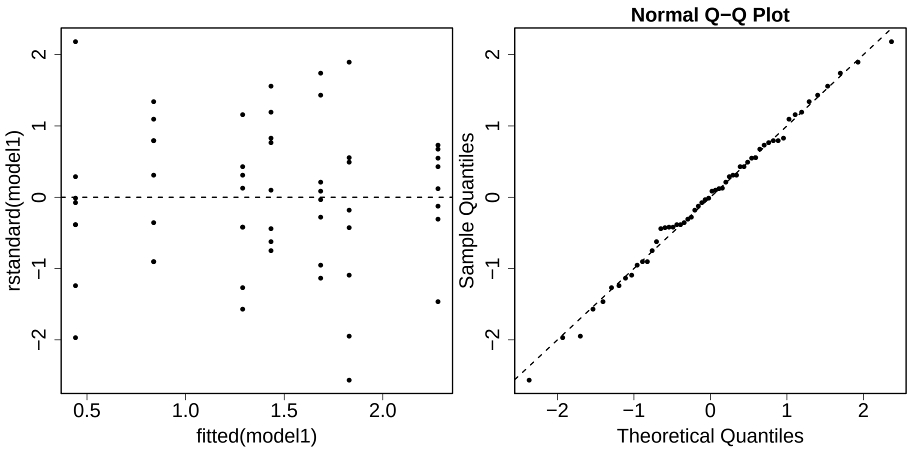
Man kan også argumentere udfra resultatet i spørgsmål 1: Fordelingen af Y er singulær, dvs. $Y_2 = a + bY_1$ med sandsynlighed 1 for passende værdier a og b. Udfra middelværdierne ser vi at $a = 0$, udfra varianserne at $b = 2$. Altså er $Y_2 = 2Y_1$, eller $X_3 = X_1 - X_2$.

Opgave 3

1. Modellen er en multipel regressionsmodel og fittes med kommandoen

```
model1 <- lm(styrke ~ A + B + C, data=limData)
```

Residualplot og QQ-plot for de standardiserede residualer er vist nedenfor:



Begge plots ser yderst fornuftige ud! I residualplottet ligger værdierne cirka symmetrisk om nul og med cirka samme spredning henover x-aksen. I QQ-plottet ligger punkterne nydeligt omkring 0/1 linien.

Bemærkning: Alle tre prædiktorer har kun to mulige værdier (0/1), og den multiple regressionsmodel er derfor sammenfaldende med den additive tresidede variansanalysemodel fra Stat2.

2. Estimaterne er følgende:

$$\hat{\alpha} = 0.4425,\ \hat{\beta}_1 = 0.9906,\ \hat{\beta}_1 = 0.8469,\ \hat{\beta}_1 = 0.3963,\ \hat{\sigma}^2 = 0.1705^2 = 0.0291.$$

Parameteren α er den forventede styrke uden tilsætning af nogen af de tre komponenter. Parametrene $\beta_1, \beta_2, \beta_3$ er den forventede ændring i styrke når A, B hhv. C tilsættes. Endelig er σ spredningen i fordelingen, dvs. udtryk for den "typiske"afvigelse fra middelværdien.

3. Den prædikterede værdi er

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_3 = 1.829.$$

Prædiktionsintervallet er givet i eksempel 10.31, og kan beregnes i R vha. funktionen `predict`:

```
newData <- data.frame(A=1, B=0, C=1)
predict(model1, newData, interval="p")
```

Vi får intervallet $(1.474, 2.185)$. En ny observation med A og C, men ikke B tilsat vil med 95% sandsynlighed havne i dette interval.

4. Den interessante parameterfunktion er $\delta = \beta_1 - \beta_2$. Denne kan skrives som

$$\delta = \begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \psi^T \gamma$$

hvor definitionen af ψ og γ fremgår af opskrivningen. Vi kan nu bruge eksempel 10.30 til at bestemme de relevante størrelser:

$$\begin{aligned} \hat{\hat{\delta}} &= \psi^T \hat{\gamma} = 0.1438 \\ \mathrm{Var}(\hat{\hat{\delta}}) &= \psi^T \mathrm{Var}(\hat{\gamma}) \psi = 0.003632 \\ \mathrm{SE}(\hat{\hat{\delta}}) &= 0.0603. \end{aligned}$$

Vi kan bruge `vcov` til at finde variansmatricen i R. Det tilhørende 95% konfidensinterval er

$$0.1438 \pm 2.007 \cdot 0.0603 = (0.0228\,,\, 0.2647)$$

hvor vi har benyttet at 97.5% fraktilen i t_{52} fordelingen er 2.007. Konfidensintervallet indeholder ikke nul, så der er tegn på at komponent A virker bedre end komponent B.

Alternativt kan vi teste hypotesen $H : \beta_1 = \beta_2$. Vi får

$$t = \frac{\hat{\delta}}{\mathrm{SE}(\hat{\hat{\delta}})} = 2.39$$

der skal vurderes in t_{52} fordelingen. Dette giver p-værdien 0.021, så hypotesen forkastes og konklusionen er (naturligvis) som før.

5. Synergieffekten svarer til at middelværdien har formen

$$\mathrm{E}Y_i = \alpha + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i + \varphi A_i B_i, \quad i = 1, \ldots, 56$$

hvor det sidste led jo netop er 1 hvis både A og B er tilsat. Modellen fittes fx som følger, hvor vi først laver produktvariablen `AB`:

```
limData <- transform(limData, AB=A*B)
model2 <- lm(styrke ~ A + B + C + AB, data=limData)
```

Synergiparameteren estimeres til $\hat{\varphi} = 0.0333$ med 95% konfidensinterval $(-0.1660\,,\, 0.2326)$. Hypotesen $H : \varphi = 0$ kan testes med et t-test hvor man får $t = 0.336$ og $p = 0.74$. Der er altså ikke evidens for synergi.