Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1, vejledende besvarelse

30. juni 2016

Dette er en vejledende besvarelse. Se og kør evt. også R-programmet `juni16.R`.

Opgave 1

1. På grund af uafhængighed er likelihoodfunktionen givet ved

$$L_x(\beta) = \prod \frac{(\beta t_i)^{x_i}}{x_i!} e^{-\beta t_i} \propto \beta^{S_x} e^{-\beta S_t}, \quad \beta > 0$$

hvor S_x og S_t er summerne fra opgaveteksten. Minus-log-likelihoodfunktionen, scorefunktion og informationsfunktion bliver derfor

$$\ell_x(\beta) = -\log L_x(\beta) = -S_x \log \beta + \beta S_t, \quad \beta > 0$$
$$S_x(\beta) = \ell_x'(\beta) = -\frac{S_x}{\beta} + S_t, \quad \beta > 0$$
$$I_x(\beta) = \ell_x''(\beta) = \frac{S_x}{\beta^2}, \quad \beta > 0$$

Notationen er faktisk temmelig uheldig her, eftersom S_x både dækker over scorefunktionen og summern af x'erne.

Endelig fås Fisherinformationen

$$i(\beta) = \mathrm{E}_\beta I_X(\beta) = \frac{\mathrm{E}_\beta S_X}{\beta^2} = \frac{\mathrm{E}_\beta \sum X_i}{\beta^2} = \frac{\sum \beta t_i}{\beta^2} = \frac{\sum t_i}{\beta} = \frac{S_t}{\beta}.$$

2. Vi betragter først estimatoren

$$\tilde{\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{t_i} \quad \text{og}$$

Vi bestemmer middelværdi og varians:

$$\mathrm{E}_\beta \tilde{\beta} = \frac{1}{n} \sum \frac{\beta t_i}{t_i} = \beta$$
$$\mathrm{V}_\beta \tilde{\beta} = \frac{1}{n^2} \sum \mathrm{V}_\beta \left(\frac{X_i}{t_i} \right) = \frac{1}{n^2} \sum \frac{\beta t_i}{t_i^2} = \frac{\beta}{n^2} \sum \frac{1}{t_i}$$

Da $\mathrm{E}_\beta \tilde{\beta} = \beta$ for alle β, er $\tilde{\beta}$ central en central estimator for β.

Men $\mathrm{V}_\beta \tilde{\beta} \neq i^{-1}(\beta)$, så den nedre grænse for variansen for en central estimator fra Cramer-Rao opnås ikke. Vi kan strengt taget (endnu) ikke udelukke at estimatoren $\tilde{\beta}$ har den mindste varians blandt centrale estimatorer, men Cramer-Rao fortæller os det ikke.

Vi betragter så estimatoren

$$\check{\beta} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} t_i}.$$

og bestemmer middelværdi og varians:

$$\mathrm{E}_\beta \check{\beta} = \frac{\sum \beta t_i}{\sum t_i} = \beta$$
$$\mathrm{V}_\beta \check{\beta} = \frac{\sum \beta t_i}{(\sum t_i)^2} = \frac{\beta}{\sum t_i} = \frac{\beta}{S_t} = i^{-1}(\beta)$$

Vi ser at $\check{\beta}$ er central og variansminimal (jf. Cramer-Rao). Vi kan således nu også slutte at $\tilde{\beta}$ *ikke* er variansminimal.

3. Betragt $(x_1, \ldots, x_n)$ med $S_x \neq 0$. Vi løser først scoreligningen:

$$S_x(\beta) = 0 \Leftrightarrow \frac{S_x}{\beta} = S_t \Leftrightarrow \beta = \frac{S_x}{S_t}$$

Da $I_x(\beta) > 0$ for alle β er løsningen til scoreligningen faktisk et globalt minimum for ℓ_x hvis $S_x \neq 0$. Desuden er løsningen positiv når $S_x > 0$.

Hvis summen $S_x =$, er alle $x_i = 0$. Så er

$$L_x(\beta) = \prod e^{-\beta t_i} = e^{-\beta S_t}$$

der ikke har maksimum på $(0, \infty)$. Men L_x er aftagende så det er naturligt at udvide parametermængden og sætte $\hat{\beta} = 0$.

Konklusion: ML estimatoren defineres naturligt som

$$\hat{\beta} = \frac{S_X}{S_t} = \check{\beta}.$$

4. For de givne data er MLE $\hat{\beta} = \check{\beta} = 3.818$ med standard error

$$\mathrm{SE}(\hat{\beta}) = \sqrt{\frac{\hat{\beta}}{\sum t_i}} = 0.833,$$

så Wald 95% konfidensintervallet er

$$\hat{\beta} \pm 1.96 \cdot \mathrm{SD}(\hat{\beta}) = 3.818 \pm 1.96 \cdot 0.833 = (2.185\,,\, 5.451)$$

Betragt så hypotsen $H : \beta = 2$. Likelihood ratio teststørrelsen er

$$LR(x) = 2\Big(\ell_x(2) - \ell_x(\hat{\beta})\Big) = 2\Big(-S_x \log 2 + 2S_t + S_x \log \hat{\beta} - \hat{\beta} S_t\Big) = 7.16$$

Denne skal vurderes i χ^2 fordelingen med 1 frihedsgrad:

$$p = P\Big(LR(X) \geq\geq 7.16\Big) = 0.007$$

så hypotesen forkastes. Der er altså evidens i data på at β er større end 2.

5. Jeg fik følgende skema:

			Teoretiske værdier		Simulation	
Estimator	n	β	middelværdi	spredning	gennemsnit	spredning
$\tilde{\beta}$	10	3	3	0.937	2.969	0.924
	20	3	3	0.735	2.981	0.730
$\check{\beta}$	10	3	3	0.739	2.965	0.738
	20	3	3	0.534	2.991	0.527

Vi ser at de teoretiske og simulerede værdier stemmer pænt overens. Specielt er begge estimatorer centrale og $\hat{\beta} = \check{\beta}$ har mindre spredning end $\tilde{\beta}$. Desuden bliver spredningerne mindre når sample size n vokser. Med andre ord: Alt er som forventet.

Opgave 2

1. X er normalfordelt på $\mathbb{R}^2$ med middelværdi 0 og varians

$$\mathrm{V}X = \begin{pmatrix} 1 & 3 & 0 & 0 \\ 3 & 9 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 3 & 9 \end{pmatrix}.$$

   Variansmatricen har determinant nul og er derfor singulær. Fordelingen af X er således en singulær normalfordeling på $\mathbb{R}^4$.

2. X_2 og X_4 er begge $N(0,9)$-fordelte, så de har samme fordeling.

   Alligevel har $X_1 + X_2$ og $X_1 + X_4$ forskellige fordeling. Dette skyldes at X_1 og X_2 er afhængige — faktisk er $X_2 = 3X_1$ — mens X_1 og X_4 er uafhængige.

   Mere specifikt kan vi finde den simultane fordeling af $X_1 + X_2$ og $X_1 + X_4$:

$$\begin{pmatrix} X_1 + X_2 \\ X_1 + X_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}$$

   der er normalfordelt med middelværdi 0 og varians

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 & 0 & 0 \\ 3 & 9 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 3 & 9 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 16 & 4 \\ 4 & 10 \end{pmatrix}.$$

   Altså er $X_1 + X_2 \sim N(0,16)$, mens $X_1 + X_4 \sim N(0,10)$.

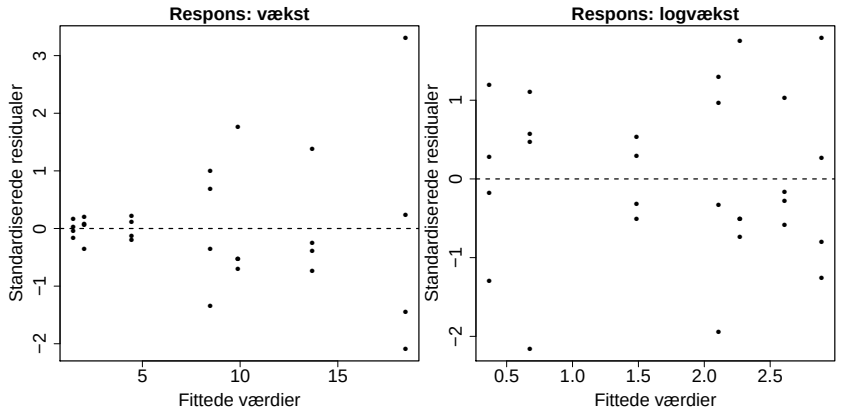
   Eftersom denne variansmatric er regulær (determinanten er 144), er den simultane fordeling af $X_1 + X_2$ og $X_1 + X_4$ en regulær normalfordeling og „fylder derfor hele $\mathbb{R}^2$“. Der findes således ikke et underrum U som beskrevet i spørgsmålet.

Opgave 3

1. Modellerne fittes med kommandoerne

```
m0 <- lm(vaekst ~ dosisGrp, data=rotteData)
m1 <- lm(logvaekst ~ dosisGrp, data=rotteData)
```

   Residualplottene ses nedenfor. Der er oplagt problemer med varianshomogenitet når `vaekst` bruges som respons: variansen vokser når de fittede værdier vokser. Når `logvaekst` benyttes som respons, ser variansen derimod ud til at være konstant.



2. Forskellen i EX mellem dosis 3 og dosis 0 aflæses direkte fra `summary` af model `m1`:

$$\text{Estimat: } 2.522, \quad 95\% \text{ KI: } (2.195\,,\, 2.849)$$

Forskellen i EX mellem dosis 3 og dosis 2 kan fx fås ved at genfitte modellen med dosisgruppe `d2` som reference. Man får:

$$\text{Estimat: } 0.6196, \quad 95\% \text{ KI: } (0.2923\,,\, 0.9468)$$

3. Den lineære regressionsmodel fittes med kommandoen

```
m2 <- lm(logvaekst ~ dosis, data=rotteData)
```

Forskellen i forventet log-vaekst ved dosis 3 og dosis 0 er 3β hvor β er hældningen i modellen. Vi får:

$$\text{Estimat: } 2.617, \quad 95\% \text{ KI: } (2.295\,,\, 2.939)$$

Forventet log-vaekst ved dosis 0 er netop interceptparamereren i modellen. Aflæsning giver

$$\text{Estimat: } 0.4642\,, \quad 95\% \text{ KI: } (0.2708\,,\, 0.6575)$$

4. Den kvadratiske regressionsmodel fittes med kommandoerne

```
rotteData$dosisKvd <- rotteData$dosis^2
m3 <- lm(logvaekst ~ dosis + dosisKvd, data=rotteData)
```

Modellen siger at

$$\mathrm{E}X = \beta_0 + \beta_1 d + \beta_2 d^2$$

så den ønskede forskel er

$$\delta = \beta_0 + 3\beta_1 + 9\beta_2 - \beta_0 = 3\beta_1 + 9\beta_2$$

der kan skrives som $C\beta$ hvor $C = \begin{pmatrix} 0 & 3 & 9 \end{pmatrix}$.

Estimatet for δ er $\hat{\delta} = C\hat{\beta} = 2.617$ og den tilhørende standard error er

$$\mathrm{SE}(\hat{\delta}) = \sqrt{C\,\mathrm{Var}(\hat{\beta})\,C^T} = 0.135$$

Konfidensintervallet bliver således

$$\hat{\delta} \pm t_{0.975,26} \cdot \mathrm{SE}(\hat{\delta}) = (2.339\,,\, 2.896)$$

Bemærk at estimaterne i spørgsmål 3 og 4 er ens, men det er „tilfældigt“ (specielt for disse data, ikke noget strukturelt ved modellerne).

5. Der gælder

$$L_{\text{linreg}} \subset L_{\text{kvadreg}} \subset L_{\text{anova}}$$

og begge mængdeinklusioner betyder „ægte underrum“.

Modelreduktionen starter dermed i den ensidede variansanalyse, $\xi \in L_{\text{anova}}$ hvor ξ betegner middelværdivektoren.

Hypotesen $H : \xi \in L_{\text{kvadreg}}$ testes med et F-test. Vi får

$$f = 1.95, \quad p = 0.14 \quad (\text{beregnet i } F(4,21))$$

så hypotesen kan ikke afvises.

Vi bruger derfor den kvadratiske regressionsmodel. Hypotesen $H : \xi \in L_{\text{kvadreg}}$ eller $H : \beta_2 = 0$ kan testes med et t-test (eller et F-test). Vi får

$$t = -3.12, \quad p = 0.004 \quad (\text{beregnet i } t(25))$$

så hypotesen afvises.

Slutmodellen er således den kvadratiske regressionsmodel, hvor middelværdien beskrives som et andengradspolynomium i dosis.

6. Vi bruger den kvadratiske regressionsmodel. Prædiktionsintervallet for log-vækst kan beregnes vha. kommandoerne

```
> newData <- data.frame(dosis=2.25, dosisKvd=2.25^2)
> predict(m3, newData, interval="p")
```

Man får prædiktionen 2.498 og 95% prædiktionsintervallet på $(1.99, 3.01)$ for log-vækst og dermed et 95% prædiktionsinterval for vækst på $(7.3, 20.2)$. En vægtøgning på 18 gram er således ikke usædvanligt.