GRIZLI777

Homework 1

Statistische modellen en data analyse

Andreas Put – Ma1 Computerwetenschappen

27/04/2011

[Type the abstract of the document here. The abstract is typically a short summary of the contents of the document. Type the abstract of the document here. The abstract is typically a short summary of the contents of the document.]

Vraag 1

deel 1

$F_X(x) = \begin{cases} 1-(1-p)^{\lfloor x \rfloor} & Als\ x\ \geq 1 \\ 0 & Elders \end{cases}$

De dichtheidsfunctie $f_X(x)$ van $F_X(x)$ kan berekend worden door F_x af te leiden naar x

$$f_X(x) = \frac{\partial Fx(x)}{\partial x}$$

$$= -\ln(1-p)\ (1\ -\ \mathrm{p})^{\lfloor x \rfloor}$$

Via de formule MLE = $\mathrm{argmax}_{\theta \in \Theta} \prod_1^n f\theta(xi)$ kunnen we de maximum likelyhood estimator vinden

$\mathrm{MLE}_p = \mathrm{argmax} \prod_1^n (-\ln(1\ -\ \mathrm{p})(1\ -\ \mathrm{p})^{\lfloor x \rfloor})$

Vraag 2

Deel 1

De onderstaande formules worden veelvuldig gebruikt voor deze vraag:

$$E[X_2|X_1=x_1] = \mu_2 + \Sigma_{21}\ \Sigma_{11}^{-1}\ (x_1 - \mu_1) \quad (1) \qquad \mu = \begin{bmatrix} \mu 1 \\ \mu 2 \end{bmatrix}$$

$$Cov[X_2|X_1 = x_1] = \Sigma_{22} - \Sigma_{21}\ \Sigma_{21}^{-1}\ \Sigma_{12} \quad (2) \qquad \Sigma = \begin{bmatrix} \Sigma 11 & \Sigma 12 \\ \Sigma 21 & \Sigma 22 \end{bmatrix}$$

We beginnen bij Y|Z

Y|Z is normaal verdeeld: $\quad Y|Z \sim N(1+Z, 1) \quad (3)$

Vullen we $x_1 = \mu 1$ in in (1) dan bekomen we

$E[X_2|X_1 = \mu 1] = \mu 2 = 1+Z$ (3). Omdat Z = x1 = µ1 = 0 ➔ µ2 = 1

Vullen we x1 = 2 in en daar $\Sigma_{11} = 1$ bekomen we

$E[X_2|X_1=2] = 1+Z = 3 = 1 + \Sigma_{21}(2-0)$ ➔ $\Sigma_{21} = 1 = \Sigma_{12}$

Σ_{22} verkrijgen we via (2)

$Cov[X_2|X_1 = 0] = 1 = \Sigma_{22} - 1/1*1$ ➔ $\Sigma_{22} = 2$

⇨ $[Z\ Y]^T \sim \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \right)$

Vervolgens analyseren we X|Y,Z

Vullen we $x_1 = \mu1 = [0\ 1]^T$ in in (1) dan bekomen we $E[X_2|X_1 = \mu1] = \mu2 = 1-Y = 0$

Vullen we $x_1 = [1,1]^T$ in en daar $\Sigma_{11} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$ is, bekomen we

$E[X_2|X_1 = x_1] = 1-Y = 0 = \Sigma_{21} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \Sigma_{21} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = 0$ (4)

Met $x_1 = [0\ 0]^T$ bekomen we $E[X_2|X_1 = x_1] = 1-Y = 1 = \Sigma_{21} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \Sigma_{21} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = 1$ (5)

Uit (4) en (5) kunnen we Σ_{21} halen ➔ $\Sigma_{21} = [1\ \ 2] = \Sigma_{12}^T$

Σ_{22} verkrijgen we via (2)

$Cov[X_2|X_1 = \mu1] = 1 = \Sigma_{22} - \Sigma_{12}\, \Sigma_{11}^{-1}\, \Sigma_{12}^T$ ➔ $\Sigma_{22} = 3$

⇨ $[Z\ \ Y\ \ X]^T \sim \left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \right)$ ➔ $[X\ \ Y\ \ Z]^T \sim \left(\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \right)$

Y|X,Z

Hiervoor is de verdeling van $[X,Z]^T$ nodig:

$[X\ Z]^T \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \right)$ $\quad$ $X,Z,Y \sim \left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 2 \end{bmatrix} \right)$

$E[Y|X,Z = \mu] = 1 + [2\ \ 1] \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} X - 0 \\ Z - 0 \end{bmatrix} \right) = ½(Z+X)$

$Cov[Y|X,Z = \mu] = 2 - [2\ \ 1] \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 1 \end{bmatrix}) = 2 - 1.5 = 0.5$

$Y|X,Z \sim (½(Z+X), 0.5)$

Deel 2

De verdeling van $[U\ \ V]^T$ waar U = 1 + Z en V = 1 - Y :

$U = 1 + Z$ ➔ $\mu_u = 1 + \mu_z$; $S_u = S_z$ ➔ $U \sim (1,1)$ - Er verandert niets aan de variantie na optelling van een constante.

$V = 1 - Y$ ➔ $\mu_v = 1 - \mu_y$; $S_v = S_y$ ➔ $V \sim (0,2)$

$\mu_{U,V} = [\ 1\ \ \ 0\ ]$

$\Sigma_{U,V} = \begin{bmatrix} \Sigma z & \Sigma zy \\ \Sigma zy & \Sigma y \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

Deel 3

Omdat de variantie van U gelijk is aan die van Z, en de variantie van V is gelijk aan de variantie van Y, is de covariantiematrix van [U,Y] gelijk aan die van [U,V]

$E(Y|U = 2) = \mu_Y + \Sigma_{21}\, \Sigma_{11}^{-1}\, (2 - \mu_U)$ $\quad$ $\begin{bmatrix} U \\ Y \end{bmatrix} \sim \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \right)$

$= 1 + 1/1 * (2-1) = 2$

Vraag 3

Gegeven de steekproef met verdeling

$$N_3 \sim \begin{pmatrix} \mu 1 & \\ \mu 2 & \Sigma \\ \mu 3 & \end{pmatrix}$$

De hypothese $H_0 : \mu 1 = \mu 2 = \mu 3$ v.s. $H_1 : \mu_i \neq \mu_j$ kan als volgt geëvalueerd worden.
We stellen een alternatieve hypothese op:

$$H_0^{'} : \mu = \begin{bmatrix} \mu 1 - \mu 1 \\ \mu 2 - \mu 1 \\ \mu 3 - \mu 1 \end{bmatrix} = 0 \text{ v.s. } H_1' : \mu \neq 0$$

Via de T^2-test van Hotelling kan deze hypothese aangenomen of verworpen worden.
Omdat deze test affien invariant is, zal de volgende hypothese

$$H_{Y0} : A\mu = 0 \text{ v.s. } H_{Y1} : A\mu \neq 0 \qquad (\det(A) \neq 0)$$

hetzelfde resultaat opleveren. De kritieke waarde van de T^2 test zal immers hetzelfde zijn als bij de originele hypothese test.

Een geschikte matrix A hiervoor is: $A1 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 1 \end{bmatrix}$ $(\det(A1) = 1)$

A1*µ levert als resultaat: $\begin{bmatrix} \mu 1 - \mu 2 \\ 0 = (\mu 2 - \mu 2) \\ \mu 3 - \mu 2 \end{bmatrix}$

Een andere matrix die we kunnen opstellen is: $A2 = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix}$ $\det(A2) = 1$

A2*µ levert als resultaat: $\begin{bmatrix} \mu 1 - \mu 3 \\ \mu 2 - \mu 3 \\ 0 = (\mu 3 - \mu 3) \end{bmatrix}$

Door het toepassen van de transformaties door de matrix Ai wordt getest of $\mu 1 = \mu 2 = \mu 3$.

Vraag 4

Deel 1

De R code voor deze oefening is te vinden in bijlage ....

Omdat we (ongecorreleerde) gepaarde steekproeven behandelen, is het aantal genomen samples gelijk aan n = 10

De hypothesetest

$$H_0 : \mu_1 = \mu_2 \text{ v.s. } H_1 : \mu_1 \neq \mu_2$$

kan worden opgelost door eerst de twee gepaarde steekproeven samen te voegen. De kenmerken van de resulterende steekproef worden dan:

$$X_p = X_1 - X_2 \quad \text{en } S_p = \frac{n*S1 + n*S2}{2*n}$$

Met deze samengestelde gemiddelde en covariantiematrix kan een nieuwe hypothese worden opgesteld:

$$H'_0 : \mu_p = 0 \text{ v.s. } H_1 : \mu_1 \neq 0$$

De test of de hypothese geldt of niet wordt uitgevoerd door middel van de Hotelling test voor gepaarde steekproeven.

$$T^2 = n * x_p^T * S_p^{-1} * x_p$$

En zo kan de F waarde berekend worden:

$$F = \frac{(n-p)*T2}{p*(n-1)}$$

Als deze waarde groter is dan de critische waarde $F_{0.95,2,8}$ zal de hypothese verworpen worden.

In dit geval geldt : F =7.901235, $F_{0.95,2,8}$= 4.45897
(Alle tussenberekeningen en uitkomsten kunnen verkregen worden door het R bestand uit te voeren)

De hypotheektest leert ons dat $\mu_p \neq 0$, en dus $\mu_1 \neq \mu_2$, maar het zegt niets over welke elementen hiervoor verantwoordelijk zijn.

Deel 2

Om dit te onderzoeken gebruiken we betrouwbaarheidsintervallen voor de verschillende elementen van x_p. Wanneer 0 niet in dit betrouwbaarheidsinterval ligt, zal de nulhypothese verworpen worden. Omdat $X_p = X_1 - X_2$ zal als 0 niet in het betrouwbaarheidsinterval van X_{pi} ligt ook X1i ≠ X_{2i} zijn.

Het betrouwbaarheidsinterval voor $X_p(i)$ wordt via de volgende formule berekend:

$$[X_p(i) \pm t_{n-1,1-\alpha/2}\sqrt{Sp(i,i)/n}]$$

Voor $H_0 : \mu_{11} = \mu_{21}$ v.s. $H_1 : \mu_{11} \neq \mu_{21}$ bekomen we het volgende betrouwbaarheidsinterval

[0.2477406 3.7522594]

Omdat 0 niet in dit interval ligt, kunnen we de nulhypothese verwerpen.

Voor $H_0 : \mu_{12} = \mu_{22}$ v.s. $H_1 : \mu_{12} \neq \mu_{22}$ bekomen we het volgende betrouwbaarheidsinterval

$$[-1.752259 \quad 1.752259]$$

0 ligt in dit interval, en we nemen de hypothese dus aan.

```
######EXERCISE 4######
###Part 1###
matrix(c(3,1),ncol = 1)->x1
matrix(c(1,1),ncol = 1)->x2
matrix(c(4,-1 ,-1,2),nrow = 2, ncol=2, byrow=TRUE)->s1
matrix(c(2,-2 ,-2,4),nrow = 2, ncol=2, byrow=TRUE)->s2
#Merge the covariance matrices and means
sp <- (9*s1+9*s2)/(18);sp
xp <- x1-x2;xp

#T² Hotelling test
#calculate T²
t2<- 10*t(xp)%*%(solve(sp))%*%(xp);t2
#Calculate F value
f<-(10-2)*t2/((10-1)*2);f #f <- (n1+n2-k-1)*t2/((n1+n2-2)*k)
#Calculate critical value
crit<-qf(.95,2,8)
#regection regian
c(crit,Inf) ## if f in rejection region ==> Reject!


###Part 2###
#Confidence interval for x(1,1)
#If 0 lies in the CI, we accept the hypothesis, otherwise reject.
t <- qt(.975,9);t

cil1 <- xp[1] - t*sqrt((2*sp[1,1])/10)
cih1 <- xp[1] + t*sqrt((2*sp[1,1])/10)
c(cil1,cih1)

###Part 3###
#Confidence interval for x(2,1)
#If 0 lies in the CI, we accept the hypothesis, otherwise reject.
cil2 <- xp[2] - t*sqrt((2*sp[2,2])/10)
cih2 <- xp[2] + t*sqrt((2*sp[2,2])/10)
c(cil2,cih2)
```