# Statische Modellen & Data-analyse

**Homework 1**

Li Quan

27 april 2011

## Vraag 1

Gegeven de verdelingsfunctie (waarbij $\lfloor x \rfloor = \text{Floor}\,(x)$ en $p \in (0,1)$)

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

kan de dichtheidsfunctie $f_X(x)$ berekend worden:

$$f_X(x) = \frac{\partial F_X(x)}{\partial x} = \ln(1-p)(-(1-p)^{\lfloor x \rfloor})\frac{\partial \lfloor x \rfloor}{\partial x}.$$

De maximum likelihood schatter voor $p$ (gegeven een steekproef $x_1, \ldots, x_n$ uit $F_X$),

$$\hat{p}_{MLE} = \arg\max_p \sum_{i=1}^{n} \ln f_X(x_i)$$

kan gevonden worden door volgende vergelijking op te lossen naar $\hat{p}$:

$$0 = \sum_{i=1}^{n} \frac{\partial \ln f_X(x_i)}{\partial p}\Big|_{p=\hat{p}} = \sum_{i=1}^{n} \Big\{ -\frac{1}{\ln(1-\hat{p})}\frac{1}{1-\hat{p}} + \frac{\lfloor x_i \rfloor}{1-\hat{p}} \Big\}$$

Uiteindelijk vinden we (na bevestiging dat het om een maximum gaat)

$$\hat{p}_{MLE} = 1 - \exp\Big(\frac{-n}{\sum_{i=1}^{n} \lfloor x_i \rfloor}\Big).$$

Om te illustreren dat $\hat{p}_{MLE}$ een vertekende schatter is van $p$ werd dit in R gesimuleerd[1] (zie code in Appendix A) voor $n = 50$. Tabel 1 toont de resultaten hiervan: hieruit blijkt dat de MLE waarschijnlijk een onderschatter is.

[1] Voor het genereren van de steekproefrealisatie werd gebruik gemaakt van de *inverse transformation sampling*-techniek (zie bijvoorbeeld `http://en.wikipedia.org/wiki/Inverse_transform_sampling`).

| | $\hat{p}_{MLE}^{(k)}$ | | | | |
|---|---|---|---|---|---|
| $p$ | 1 | 2 | 3 | 4 | 5 |
| 0.1 | 0.097 | 0.104 | 0.108 | 0.093 | 0.090 |
| 0.3 | 0.315 | 0.323 | 0.295 | 0.300 | 0.300 |
| 0.5 | 0.473 | 0.482 | 0.531 | 0.461 | 0.553 |
| 0.7 | 0.671 | 0.655 | 0.647 | 0.640 | 0.654 |
| 0.9 | 0.854 | 0.770 | 0.875 | 0.771 | 0.780 |

**Tabel 1:** $\hat{p}_{MLE}$ voor verschillende waarden van $p$ met steekproefgrootte $n = 50$ en $k = 5$ runs (zie R-code Appendix A).

Om dit te bewijzen definiëren we eerst de bias $b(\hat{p}) = \mathbb{E}[\hat{p}] - p$. Dan moeten we aantonen dat $b(\hat{p}) \neq 0$. Algemeen werd reeds aangetoond dat MLE schatters onvertekend zijn tot op orde $n^{-1/2}$, maar vertekend op orde $n^{-1}$ [2, Eq. (20)].

Een algemene uitdrukking voor de bias van een MLE schatter $\hat{\theta}$ is [3]:

$$b(\theta) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \dots$$

Asymptotisch is de MLE dus onvertekend (dit is de asymptotische normaliteitseigenschap van MLE schatters), aangezien $\lim_{n\to\infty} b = 0$.

## Vraag 2

We beschouwen eerst $Y$ en $Z$. Dan is $\begin{pmatrix} Y \\ Z \end{pmatrix} \sim \mathcal{N}_2(\begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_{YY} & \sigma_{YZ} \\ \sigma_{ZY} & \sigma_{ZZ} \end{pmatrix})$, waarbij we volgende vergelijkingen moeten oplossen ($\mu_Z = 0$ en $\sigma_Z^2 = \sigma_{ZZ} = 1$):

$$\begin{aligned} \mathbb{E}[Y|Z = z] &= 1 + z \\ &= \mu_{YY} + \sigma_{YZ}\sigma_Z^{-1}(z - \mu_Z) = \mu_Y + \sigma_{YZ} z \\ \mathrm{Cov}[Y|Z = z] &= 1 \\ &= \sigma_{YY} - \sigma_{YZ}\sigma_Z^{-1}\sigma_{ZY} = \sigma_{YY} - \sigma_{YZ}\sigma_{ZY} \end{aligned}$$

De oplossing hiervoor is $\mu_Y = 1$, $\sigma_{YZ} = \sigma_{ZY} = 1$ en $\sigma_{YY} = 2$.

Dus dan is[2] $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}_3(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix})$, met bijkomende vergelijkingen:

$$\begin{aligned}
\mathbb{E}[X|Y = y, Z = z] &= 1 - y \\
&= \mu_1 + \begin{pmatrix} \sigma_{12} & \sigma_{13} \end{pmatrix} \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{pmatrix}^{-1} \begin{pmatrix} y - \mu_2 \\ z - \mu_3 \end{pmatrix} \\
\mathrm{Cov}[X|Y = y, Z = z] &= 1 \\
&= \sigma_{11} - \begin{pmatrix} \sigma_{12} & \sigma_{13} \end{pmatrix} \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{21} \\ \sigma_{31} \end{pmatrix}
\end{aligned}$$

De oplossingen van het volledige stelsel zijn dan: $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 0$, $\sigma_{11} = 3$, $\sigma_{22} = 2$, $\sigma_{33} = 1$, $\sigma_{12} = \sigma_{21} = 2$, $\sigma_{13} = \sigma_{31} = 1$ en $\sigma_{23} = \sigma_{32} = 1$. Dus is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}_3(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}).$$

Na herschikken tot $\begin{pmatrix} Y \\ X \\ Z \end{pmatrix} \sim \mathcal{N}_3(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 2 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix})$ kunnen we gemakkelijk $Y|X, Z \sim \mathcal{N}(\mu_Y, \sigma_{YY})$ berekenen waarbij

$$\begin{aligned}
\mu_Y &= 1 + \begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} X \\ Z \end{pmatrix} = \frac{X + Z}{2} \\
\sigma_{YY} &= 2 - \begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \frac{1}{2}
\end{aligned}$$

Voor de verdeling van $(U, V)^\tau$ zoeken we eerst de verdeling van $U$ en $V$. Hiervoor gebruiken we volgende eigenschap: als $X \sim \mathcal{N}(\mu, \sigma^2)$, dan is een lineaire transformatie $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. Uit $U = 1 + Z$ en $V = 1 - Y$ volgt dan dat $U \sim \mathcal{N}(1, 1)$ en $V \sim \mathcal{N}(0, 2)$. Verder volgt uit de lineariteitseigenschap van de covariantie dat $\mathrm{cov}(U, V) = \mathrm{cov}(1 + Z, 1 - Y) = \mathrm{cov}(Z, -Y) = -\mathrm{cov}(Z, Y) = -1$ zodat

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}_2(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}).$$

[2] Hierbij worden cijferindices gebruikt: $X \equiv 1$, $Y \equiv 2$ en $Z \equiv 3$.

Aangezien $\mathrm{cov}(U,Y) = \mathrm{cov}(1+Z,Y) = \mathrm{cov}(Z,Y) = 1$, is $\begin{pmatrix} U \\ Y \end{pmatrix} \sim \mathcal{N}_2(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix})$. Dan is $\mathbb{E}(Y|U=u) = \mu_Y + \sigma_{YU}\sigma_{UU}^{-1}(u-\mu_U)$ en dus

$$\mathbb{E}(Y|U=2) = 1 + \frac{1}{1}(2-1) = 2.$$

## Vraag 3

Definieer $\boldsymbol{X}_i' = \begin{pmatrix} X_{i1} \\ X_{i2} \\ X_{i3} \end{pmatrix} - \begin{pmatrix} X_{i1} \\ X_{i1} \\ X_{i1} \end{pmatrix}$. Dan is $\mathbb{E}[\boldsymbol{X}_i'] = \boldsymbol{\mu}' = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \end{pmatrix}$. Dan is de hypothese equivalent met $\boldsymbol{A}\boldsymbol{\mu}' = \boldsymbol{\mu}_0 = \boldsymbol{0}$, waarbij $\boldsymbol{A} = \boldsymbol{I}_3$ (of eender welke niet-singuliere matrix, bijvoorbeeld $\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}$). Dus kunnen we eenvoudig de (affien invariante) $T^2$-test van Hotelling gebruiken waarbij:

$$\begin{aligned} \overline{\boldsymbol{X}}' &= \sum_{i=1}^{n} \boldsymbol{X}_i' \\ \boldsymbol{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i' - \overline{\boldsymbol{X}}')(\boldsymbol{X}_i' - \overline{\boldsymbol{X}}')^\tau \\ T^2 &= n(\overline{\boldsymbol{X}}' - \boldsymbol{\mu}_0)^\tau \boldsymbol{S}^{-1}(\overline{\boldsymbol{X}}' - \boldsymbol{\mu}_0) = n\overline{\boldsymbol{X}}'^\tau \boldsymbol{S}^{-1}\overline{\boldsymbol{X}}' \end{aligned}$$

## Vraag 4

Definieer $\boldsymbol{D}_i = \boldsymbol{X}_{i1} - \boldsymbol{X}_{i2}$, $\boldsymbol{S}_D = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2} = \frac{\boldsymbol{S}_1+\boldsymbol{S}_2}{2}$ (aangezien $n = n_1 = n_2$) en $\boldsymbol{\mu}_D = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. (Alle hypothesetesten gebeuren op het significantieniveau $\alpha = 0.05$.)

De nulhypothese $\boldsymbol{\mu_1} = \boldsymbol{\mu}_2$ is dan equivalent met $\boldsymbol{\mu}_D = \boldsymbol{0}$. We gebruiken dus de multivariate *gepaarde* $T^2$-test van Hotelling [1] waarbij $T^2 = n\overline{\boldsymbol{d}}^\tau \boldsymbol{S}_D^{-1}\overline{\boldsymbol{d}}$. Dan kan volgende $F$-statistiek gebruikt worden: $F = \frac{n-p}{p(n-1)}T^2 \sim F_{p,n-p}$.

Voor de gegeven realisatie ($n = 10$ en $p = 2$) wordt dit dus:

$$\begin{aligned} \overline{\boldsymbol{d}} &= \begin{pmatrix} 3 & 1 \end{pmatrix}^\tau - \begin{pmatrix} 1 & 1 \end{pmatrix}^\tau = \begin{pmatrix} 2 & 0 \end{pmatrix}^\tau \\ \boldsymbol{s}_D &= \frac{1}{2}\left\{\begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix} + \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}\right\} = \begin{pmatrix} 3 & -1.5 \\ -1.5 & 3 \end{pmatrix} \\ t^2 &= 10\begin{pmatrix} 2 & 0 \end{pmatrix}\begin{pmatrix} 0.444 & 0.222 \\ 0.222 & 0.444 \end{pmatrix}\begin{pmatrix} 2 & 0 \end{pmatrix}^\tau = 17.778 \\ f &= \frac{8}{18}t^2 = 7.902 \end{aligned}$$

De $p$-waarde is $p = 2P_{H_0}(F \geq |f|) = 2(1 - 0.9872) = 0.0255 < 0.05$. Dus we verwerpen de nulhypothese.

Voor de hypothesetesten voor de aparte componenten gebruiken we volgende formule voor het $100(1-\alpha)\%$ betrouwbaarheidsinterval (BI):

$$\boldsymbol{u}^\tau \overline{\boldsymbol{D}} \pm t_{n-1,1-\alpha/2}\sqrt{\frac{\boldsymbol{u}^\tau \boldsymbol{S}_D \boldsymbol{u}}{n}}.$$

Voor de nulhypothese $H_0 : \mu_{11} = \mu_{21}$ kiezen we $\boldsymbol{u} = \begin{pmatrix} 1 & 0 \end{pmatrix}^\tau$. We krijgen dan als BI

$$\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \end{pmatrix} \pm t_{9,0.975}\sqrt{\frac{\begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 3 & -1.5 \\ -1.5 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix}^\tau}{10}}.$$

Met $t_{9,0.975} = 2.262$ geeft dit dan als BI $= [0.761; 3.239]$. We verwerpen dus $H_0$ aangezien het BI niet 0 bevat.

Analoog voor de nulhypothese $H_0 : \mu_{12} = \mu_{22}$ (waarbij $\boldsymbol{u} = \begin{pmatrix} 0 & 1 \end{pmatrix}^\tau$), verkrijgen we uiteindelijk het BI $= [-1.239; 1.239]$. Hier aanvaarden we dus wel $H_0$.

## Referenties


[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, second edition, Sept. 1984.

[2] D. R. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):pp. 248–275, 1968.

[3] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, Mar. 1993.

# A. R-code

```
#Li Quan - April 27, 2011
#homework 1: question 1, part 2


#the given cdf function
mycdf <- function(p, x) {
 ifelse(x >= 1, 1 - (1-p)**(floor(x)),  0)
}

#set the p in (0,1)
p <- 0.9
cdf <- function(x) mycdf(p, x)

#inverts the cdf (using numeric solver)
# http://en.wikipedia.org/wiki/Inverse_transform_sampling
# http://stackoverflow.com/questions/1594121/r-how-do-i-best-simulate-an-arbitrary
    -univariate-random-variate-using-its-probab
inversefun.cdf<-function(x, cdf, starting.value=0, precision=1e-6){
 #we first search a lower bound
 low.found<-FALSE
 low<-starting.value
 while(!low.found){
  if(cdf(low)>=(x-precision))
   low<-low-(low-starting.value)^2-1
  else
   low.found<-TRUE
 }
 #... and an upper bound
 up.found<-FALSE
 up<-starting.value
 while(!up.found){
  if(cdf(up)<=(x+precision))
   up<-up+(up-starting.value)^2+1
  else
   up.found<-TRUE
 }
 # solve this equation
 uniroot(function(y) cdf(y)-x, c(low, up))$root
}

#sample size
n <- 50
#generates n random variables of the distribution using inverse transform sampling
vars<- sapply(runif(n), function(x) inversefun.cdf(x,cdf))
hist(vars)

#calculates the estimator for p
#we immediately use the explicit formulation for p_MLE
calcEstimator<-function(vars) {
  n <- length(vars)
  vars_floors <- sapply(vars, 'floor') #we apply the floor function on each
      element of the list
  result <- 1 - exp(-n/sum(vars_floors));result
}

pMLE <- calcEstimator(vars)
#print some information, we could use hypothese tests but we leave it up to the
    reader
cat("Sample size n was", n, "\nThe original p was", p, "\nThe mle for p is", pMLE)
```