PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
Escuela de Posgrado / Maestría en Estadística /
Estadística Computacional

Examen Parcial - Parte 2

Anthony Enrique Huertas Quispe Cod:20173728
Prof: Cristian Bayes

Pregunta 4

Una variable aleatoria X, definida en toda la recta, tiene distribución normal asimétrica estándar (Azzalini, 1985) con parámetro de asimetría λ, esto es $X \sim SN(\lambda)$, si su función de densidad es dad por la siguiente expresión:

$$f(x) = 2\phi(x)\Phi(\lambda x) \tag{1}$$

Propiedad: Sean $Z_1 \sim N(0,1)$ y $Z_2 \sim N(0,1)$ independientes, entonces

$$X = \frac{1}{\sqrt{1+\lambda^2}} Z_1 + \frac{\lambda}{\sqrt{1+\lambda^2}} |Z_2| \sim SN(\lambda) \tag{2}$$

a) Encuentre el estimador de momentos para λ en este modelo.

Solución. El método de momentos, nos establece que el j-ésimo momento (poblacional) de X está dado por

$$m_j \quad = \quad E[X^j] \tag{3}$$

La estimación por dicho método nos dice que teniendo una muestra aleatoria $X_1, \ldots, X_n$ de X, cuya distribución depende de los parámetros $\theta_1, \ldots, \theta_k$, entonces sus respetivos estimadores corresponden a la solución del siguiente sistema de k ecuaciones:

$$m_j \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i^j, \;\; j = 1, \ldots, k \tag{4}$$

Dado que la distribución de X solo depende de λ, entonces el estimador de este parámetro se determina mediante:

$$E[X] \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{x} \tag{5}$$

Calculemos $E[X]$ usando la propiedad (2):

$$E[X] = \frac{1}{\sqrt{1+\lambda^2}} \underbrace{E[Z_1]}_{0} + \frac{\lambda}{\sqrt{1+\lambda^2}} E[|Z_2|] = \frac{\lambda}{\sqrt{1+\lambda^2}} E[|Z_2|] \tag{6}$$

Sabemos que $|Z_2|$ tiene fdp definida como:

$$f_{|Z_2|}(y) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}, \; y > 0 \tag{7}$$

Entonces $E[|Z_2|]$ se calcula de la siguiente forma

$$E[|Z_2|] = \int_0^\infty y \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = \sqrt{\frac{2}{\pi}} \int_0^\infty y e^{-\frac{1}{2}y^2} dy = \sqrt{\frac{2}{\pi}} \left(-e^{-\frac{1}{2}y^2} \Big[_0^\infty \right) = \sqrt{\frac{2}{\pi}} \tag{8}$$

Reemplazamos (8) en (6),

$$E[X] \quad = \quad \frac{\lambda}{\sqrt{1+\lambda^2}} \sqrt{\frac{2}{\pi}} \tag{9}$$

Usando (9) en (5), determinamos el estimador de momentos para λ:

$$\frac{\lambda}{\sqrt{1+\lambda^2}}\sqrt{\frac{2}{\pi}} = \bar{x} \quad \Rightarrow \quad \lambda = \bar{x}\sqrt{\frac{1}{\frac{2}{\pi} - \bar{x}^2}} \tag{10}$$

□

b) Indique en que casos el estimador existe.

Solución. Se necesitará que se cumpla el criterio de la raiz de acuerdo a lo establecido en (10):

$$0 < \frac{1}{\frac{2}{\pi} - \bar{x}^2} \Rightarrow 0 < \frac{2}{\pi} - \bar{x}^2 \Rightarrow \bar{x}^2 < \frac{2}{\pi} \Rightarrow 0 \leq |\bar{x}| < \sqrt{\frac{2}{\pi}} \tag{11}$$

Por tanto, para que exista λ debe satisfacerse (11). □

c) Realice un estudio de simulacion para diferentes valores de $\lambda = 0, 5, 10$ y tamaños de muestra $n = 20, 50, 100$ para calcular la probabilidad de que el estimador de momentos exista, además estudie el comportamiento del estimador utilizando el sesgo y el error medio cuadrático.

Solución. Se llevará a cabo un estudio con 10000 simulaciones para calcular la probabilidad de que el estimador de λ exista.

Listing 1: Definiciones iniciales.

```
# Numero de Simulaciones
M=10000

#Tamanos de Muestra
n=c(20,50,100)

# Parametros
lam=c(0,5,10)

# Matriz de resultados
T<-matrix(0,M,3*3)
colnames(T)<-c("lam0.20","lam0.50","lam0.100",
               "lam5.20","lam5.50","lam5.100",
               "lam10.20","lam10.50","lam10.100")
```

Listing 2: Calculando la probabilidad.

```
#Estudio de simulacion
for(l in 1:3){
        for(h in 1:3){
                for(j in 1:M){
                        z1=rnorm(n[h])
                        z2=abs(rnorm(n[h]))
                        x=z1/sqrt(1+lam[l]**2)+lam[l]*z2/sqrt(1+lam[l]**2)
                        if((mean(x))**2<(2/pi)){r=1}else{r=0}
                        T[j,h+3*(l-1)]=c(r)
                }
        }
}
PROBABILIDAD=colMeans(T)
```

TABLA DE PROBABILIDADES

n	$\lambda = 0$	$\lambda = 5$	$\lambda = 10$
20	0.9999	0.5571	0.5245
50	1	0.5715	0.5315
100	1	0.5913	0.5349

Se han calculado, para diferentes valores de λ, las probabilidades de que existan sus estimadores con tamaños de muestra de 20,50 y 100.
Se observa que la probabilidad de que exista el estimador de $\lambda = 0$(distribución simétrica) es prácticamente 1; a diferencia de $\lambda = 5, 10$ que son cercanas a 0.5.

Se desarrollará un estudio de simluación de la variable utilizando el sesgo y el ECM

Listing 3: Estudio de simulación.

```
library(sqldf)

for(l in 1:3){
        for(h in 1:3){
                for(j in 1:M){
                        # Generando X~SN(lambda)
                        z1=rnorm(n[h])
                        z2=abs(rnorm(n[h]))
                        x=z1/sqrt(1+lam[l]**2)+lam[l]*z2/sqrt(1+lam[l]**2)
                        # Condicion de existencia del estimador
                        if((mean(x))**2<(2/pi)){
                                lambda=mean(x)*sqrt(1/((2/pi)-mean(x)**2))
                                T[j,h+3*(l-1)]=c(lambda)
                        }else{ T[j,h+3*(l-1)]=c(pi) }
                }
        }
}

# Extrayendo los estimadores existentes de la simulacion
w=list()
for(i in 1:9){
        w[[i]]=subset(T[,i],T[,i]!=pi)
}
```

Listing 4: Análisis de resultados.

```
# Calculando el Sesgo y el ECM para los estimadores
Sesgo=matrix(0,3,3)
ECM=matrix(0,3,3)

colnames(Sesgo)=c("lam0","lam5","lam10")
rownames(Sesgo)=c("n20","n50","n100")
colnames(ECM)=c("lam0","lam5","lam10")
rownames(ECM)=c("n20","n50","n100")

for(i in 1:3){
        for(j in 1:3){
                Sesgo[i,j]=mean(w[[i+3*(j-1)]])-lam[j]
                ECM[i,j]=var(w[[i+3*(j-1)]])+(Sesgo[i,j])**2
        }
}
```

Resultado:

SESGO

n	$\lambda = 0$	$\lambda = 5$	$\lambda = 10$
20	-0.001437436	-2.446166	-7.223839
50	0.002002587	-1.798376	-6.575642
100	-0.002186542	-1.072391	-5.804743

ECM

n	$\lambda = 0$	$\lambda = 5$	$\lambda = 10$
20	0.12449777	20.72604	72.23048
50	0.03611686	14.46388	60.28583
100	0.01633196	29.62848	64.08511

Logramos observar que el sesgo respecto a los estimadores de los diferentes valores de λ tienden a disminuir conforme aumenta el tamaño de muestra.
Sin embargo, con respecto al ECM, notamos claramente que disminuye para el estimador de $\lambda = 0$ conforme aumenta la muestra; mientras que con respecto a los estimadores de $\lambda = 5, 10$ no podemos decir lo mismo pues no observamos una tendencia clara. □

Pregunta 5

Se desea estudiar la relación entre temperatura y la concentración de CO_2, para esto se han recolectado las siguientes variables:

- y_i=temperatura medida como su diferencia con la temperatura actual en el periodo i-ésimo.
- x_i= la concentración de CO_2 medida en partes por millón en el periodo i-ésimo.

Se puede considerar un modelo de regresión simple dado por

$$Y \sim N(X\beta, \sigma^2 I) \tag{12}$$

donde $Y = (y_1, \ldots, y_n)^T$, $X = (1, x)$ es una matriz de dimensión $n \times 2$, 1 es un vector de unos de dimensión n, $x = (x_1, \ldots, x_n)^T$, $\beta = (\beta_1, \beta_2)^T$, $\sigma^2 > 0$ y I es la matriz de identidad.
Por otro lado, es conocido que en casos de medidas repetidas en el tiempo se puede tener una autocorrelación entre las observaciones. Por lo que un modelo alternativo sería un modelo de regresión que considere una estructura autoregresiva de primer orden entre las observaciones dado por

$$Y \sim N(X\beta, \sigma^2 C_p) \tag{13}$$

donde

$$C_p = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix} \tag{14}$$

y $0 < \rho < 1$.

a) Estime por máxima verosimilitud los modelos de regresión simple y el modelo estructura autoregresiva de primer orden.

Solución. Una v.a $Y \sim N(\mu, \Sigma)$ tiene una fdp definida por:

$$f(y) = \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-1/2} exp\left(-\frac{1}{2}(y - X\beta)^T \Sigma^{-1} (y - X\beta)\right) \tag{15}$$

Cuya función de verosimilitud, vendría dada como:

$$\begin{aligned} \mathcal{L}(\mu, \Sigma) &= f(y = y_1, \ldots, y_n \mid \mu, \Sigma) \\ &= (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - \mu)\Sigma^{-1}(y - \mu)\right) \end{aligned} \tag{16}$$

El EMV se obtiene al maximizar la función log-verosimilitud:

$$\ell(\mu, \Sigma) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(y - \mu)\Sigma^{-1}(y - \mu) \tag{17}$$

- **Modelo de regresión simple:**
  Tenemos $Y \sim N(X\beta, \sigma^2 I)$ cuyo EMV se obtiene maximizando (17):

$$\ell(\beta_1, \beta_2, \sigma) = -\frac{n}{2}log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}(y - (\beta_1 + \beta_2 x))(y - (\beta_1 + \beta_2 x)) \tag{18}$$

- **Modelo de estructura autoregresiva de primer orden:**
  Tenemos $Y \sim N(X\beta, \sigma^2 C_p)$ cuyo EMV se obtiene maximizando (17):

$$\ell(\beta_1, \beta_2, \sigma, \rho) = -\frac{n}{2}log(2\pi) - \frac{1}{2}\log(|\sigma^2 C_p|) - \frac{1}{2\sigma^2}(y - (\beta_1 + \beta_2 x))C_p^{-1}(y - (\beta_1 + \beta_2 x)) \tag{19}$$

Ahora determinaremos los EMV por cada modelo haciendo uso del método de optimización directa.

Listing 5: Extracción de datos de archivo ex1.csv.

```
base=fread("ex1.csv")
x=base$x; y=base$y; n=nrow(base)
```

Listing 6: Modelo de regresión simple, $\ell(\beta_1, \beta_2, \sigma)$.

```
#Funcion Log-verosimilitud
log.like.reg.S=function(theta){
        sigma=theta[3]
        mu=theta[1]+theta[2]*x
        0.5*n*log(2*pi)+n*log(sigma)+0.5*(t(y-mu)%*%(y-mu))*sigma**(-2)
}
```

Listing 7: Modelo autoregresivo de primer orden, $\ell(\beta_1, \beta_2, \sigma, \rho)$.

```
#Funcion Log-verosimilitud
log.like.reg.A=function(theta){
        mu=theta[1]+theta[2]*x
        sigma=theta[3]
        rho=theta[4]
        Sigma=(sigma**2)*rho^abs(matrix(rep(1:n,each=n),ncol=n)-1:n)
        0.5*n*log(2*pi)+0.5*log(det(Sigma))+0.5*(t(y-mu)%*%solve(Sigma)%*%(y-mu
            ))
}
```

Listing 8: EMV (Optimización directa).

```
# Estimadores iniciales
regre=lm(y~x)
inicial=c(regre$coefficients[1],regre$coefficients[2],sd(regre$residuals))

# EMV del Modelo de regresion simple
EMV.Simple=nlminb(inicial,log.like.reg.S)

# EMV del Modelo autoregresivo de primer orden
EMV.Auto=nlminb(c(inicial,0.5),log.like.reg.A)
```

Resultado:

	β_1	β_2	σ	ρ
Modelo de regresión Simple	-23.02414043	0.07984955	1.52899847	———
Modelo autoregresivo de primer orden	-10.86028918	0.02720078	2.18378926	0.82841750

□

b) Compare ambos modelos e indique cuál sería más apropiado en este caso. Interprete sus resultados.

Solución. Se determinará cual es el modelo más apropiado, es decir el que tenga un mejor ajuste con los datos. Para ello se establecerá una comparación usando las medidas **AIC** y BIC, en donde el mejor modelo es aquel que tenga el menor valor de AIC o BIC.

Listing 9: Comparación de modelos.

```
v1=c(EMV.Simple$par[1],EMV.Simple$par[2],EMV.Simple$par[3])
v2=c(EMV.Auto$par[1],EMV.Auto$par[2],EMV.Auto$par[3],EMV.Auto$par[4])

# Comparacion por AIC
AIC.Simple=2*log.like.reg.S(v1)+2*3
AIC.Auto=2*log.like.reg.A(v2)+2*4

#Comparacion por BIC
BIC.Simple=2*log.like.reg.S(v1)+log(n)*3
BIC.Auto=2*log.like.reg.A(v2)+log(n)*4
```

Resultado:

	AIC	BIC
Modelo de regresión simple	730.4377	720.5427
Modelo autoregresivo de primer orden	641.3112	628.118

El mejor modelo es el autoregresivo pues tanto sus valores AIC y BIC son menores que los del modelo de regresión simple. □