Estadística Computacional

Alumnos:
Córdova Proleón, Christian Therius ***Cod:*** *20173970*
Fazio Luna, Boris Manuel ***Cod:*** *20173896*
Huertas Quispe, Anthony Enrique ***Cod:*** *20173728*
Saenz Egusquiza, Miguel Angel ***Cod:*** *19921255*

Semestre: *2017-I*

Tema: Lista 2

Prof. Cristian Bayes

Pontificia Universidad Católica del Perú
Escuela de Posgrado
Maestría en Estadística

Ejercicio 1

Un estimador utilizado para reducir el efecto que puedan tener valores atípicos es la media winsorizada (winsorized mean) que es la media que se obtiene descartando el $a\,\%$ de valores extremos superiores siendo reemplazados por el cuantil $(1-a)\,\%$ y el $a\,\%$ de valores inferiores de los datos siendo reemplazados por el cuantil $a\,\%$.
Realice un estudio de simulación comparando la media winsorizada al $20\,\%$ cada lado con la media muestral. Considere 3 escenarios:

(a) datos simétricos sin valores atípicos: distribución normal.

(b) datos simétricos con valores atípicos: distribución t de Student con 4 grados de libertad.

(c) datos asimétricos: distribución gama.

En cada escenario determine los parámetros de la distibución de modo que la media y la varianza sean iguales a 1. Considere los siguientes tamaños de muestra: $n = 30$, 50 y 100. Interprete sus resultados.
Sugerencia: En R la función winsor.mean de la librería psych tiene el parámetro trim que permite calcular la media winsorizada.
Recuerde que si $X \sim t(\mu, \sigma^2, \nu)$ su función de densidad es dada por:

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{\frac{\nu+1}{2}}, \; x \in \mathbb{R}$$

donde μ es un parámetro de localización, σ un parámetro de escala y ν es un parámetro de forma que controla la curtosis de la distribución,

$$\begin{aligned} E(X) &= \mu \text{ si } \nu > 1, \\ Var(X) &= \frac{\nu}{\nu-2}\sigma^2 \text{ si } \nu > 2 \end{aligned}$$

Solución. El problema plantea verificar la eficiencia del estimador "*media winzorizada*"para reducir el efecto de los valores atípicos sobre la estimación de parámetros. En tal sentido, se medirá el desempeño de los estimadores de la "*media muestral o aritmética*" y "media winzorisada", mediante el sesgo y el error cuadrático medio. Comparándolos a su vez, frente a los siguientes factores:

- **Escenarios:**
  - Datos simétricos sin valores *outlier*: Distribución Normal, $X \sim N(\mu, \sigma^2)$, donde

    $$E[X] = \mu = 1 \text{ y } Var(X) = \sigma^2 = 1 \quad \Rightarrow \quad \mu = 1, \; \sigma = 1$$

    Función de densidad, $X \sim N(1, 1)$:

    $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}, \;\; x \in \mathbb{R}$$

  - Datos simétrico con valores *outlier*: Distribución t de Student con v grados de libertad, $X \sim t(\mu, \sigma^2, \nu)$, donde

    $$\nu = 4, \;\; E[X] = \mu = 1 \text{ y } Var(X) = \frac{4}{4-2}\sigma^2 = 1 \quad \Rightarrow \quad \begin{cases} \mu = 1 \\ \sigma = \dfrac{1}{\sqrt{2}} \\ \nu = 4 \end{cases}$$

Función de densidad, $X \sim t(1, \frac{1}{2}, 4)$:

$$f(x) \quad = \quad \frac{\Gamma(5/2)}{\Gamma(2)(2\sqrt{\pi})} \left(1 + \frac{1}{2}(x-1)^2\right)^{-5/2}, \quad x \in \mathbb{R}$$

- Datos Asimétricos: Distribución Gamma, $X \sim Gamma(\alpha, \lambda)$, donde

$$E[X] = \frac{\alpha}{\lambda} = 1 \text{ y } Var(X) = \frac{\alpha}{\lambda^2} = 1 \quad \Rightarrow \quad \alpha = 1, \ \lambda = 1$$

Función de densidad, $X \sim Gamma(\alpha, \lambda)$:

$$f(x) \quad = \quad e^{-x}, \ \ x > 0.$$

- **Tamaños de muestra:** $n = 30$, 50 y 100.

Listing 1: Definiciones iniciales.

```
library(psych)

# Numero de simulaciones
M = 10000

# Tamanos de muestra
n = c(30,50,100)

# Parametros
mu = 1; sigma = c(1,1/sqrt(2))
nu=4
alpha=1; lambda=1

# Matriz de resultados
T = matrix(0,M,2*3*3)
colnames(T) = c("MedArit.30.N","MedWinsor.30.N","MedArit.30.t","MedWinsor.30.t","
    MedArit.30.G","MedWinsor.30.G","MedArit.50.N","MedWinsor.50.N","MedArit.50.t","
    MedWinsor.50.t","MedArit.50.G","MedWinsor.50.G","MedArit.100.N","MedWinsor.100.N","
    MedArit.100.t","MedWinsor.100.t","MedArit.100.G","MedWinsor.100.G")
```

Listing 2: Estudio de simulación.

```
set.seed(100)
for(h in 1:3){
        for(j in 1:M){
                x = rnorm(n[h],mu,sigma[1])
                y = mu+sigma[2]*rt(n[h],4)
                z = rgamma(n[h],alpha,lambda)
                T[j,(6*h-5):(6*h)] = c(mean(x),winsor.mean(x,trim=0.2),mean(y),winsor.
                    mean(y,trim=0.2),mean(z),winsor.mean(z,trim=0.2))
        }
}
```

Listing 3: Análisis de resultados.

```
est = rep(c("MedArit","MedWinsor"),times=9)
dis = rep(c("N","N","t","t","G","G"),times=3)
tam = rep(n,each=4)

sesgo = colMeans(T)-mu
variancia = diag(var(T))
ecm = variancia+sesgo^2
```

Listing 4: Tabla Sesgo.

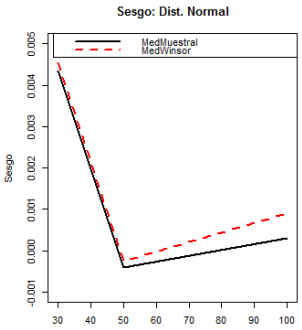
```
tab_ses = rbind(matrix(sesgo[dis=="N"],3,2,byrow=T),matrix(sesgo[dis=="t"],3,2,byrow=T
    ),matrix(sesgo[dis=="G"],3,2,byrow=T) )
tab_ses = cbind(rep(n,3),tab_ses)
colnames(tab_ses) = c("T. Muestra","MedArit","MedWinsor")
rownames(tab_ses) = rep(c("N","t","G"),each=3)
round(tab_ses,4)
```

Sesgo

Escenario	Distribución	n	Media muestral	Media winsorizada
Datos simétricos sin valores *outlier*	Normal	30	0.0043	0.0045
		50	−0.0004	−0.0003
		100	0.0003	0.0009
Datos simétricos con valores *outlier*	t-student	30	−0.0003	−0.0005
		50	0.0004	0.0009
		100	0.0002	0.0001
Datos asimétricos	Gamma	30	−0.0014	−0.1672
		50	0.0028	−0.1685
		100	−0.0003	−0.1745

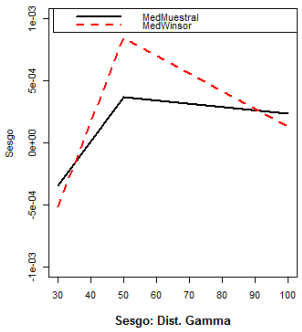
Listing 5: Gráfica Sesgo (Normal).

```
par(mfrow=c(1,3))
#Normal
plot(n,sesgo[(dis=="N")&(est=="MedArit")],ylim=c(-0.001,0.005),type=
    "l",col=1,lty=1,lwd=2,
main="Sesgo: Dist. Normal",xlab="Tamano de Muestra",ylab="Sesgo")
lines(n,sesgo[(dis=="N")&(est=="MedWinsor")],col=2,lty=2,lwd=2)
legend("topright",c("MedMuestral","MedWinsor"),col=c(1,2),lty=c(1,2)
    ,lwd=2)
```



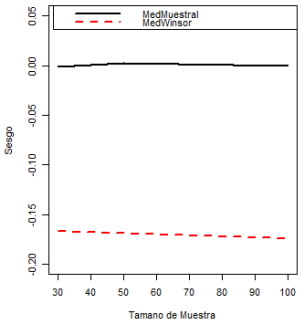
Listing 6: Gráfica Sesgo (t-Student).

```
#t-Student
plot(n,sesgo[(dis=="t")&(est=="MedArit")],ylim=c(-0.001,0.001),type=
    "l",col=1,lty=1,lwd=2,
main="Sesgo: Dist. t-Student",xlab="Tamano de Muestra",ylab="Sesgo")
lines(n,sesgo[(dis=="t")&(est=="MedWinsor")],col=2,lty=2,lwd=2)
legend("topright",c("MedMuestral","MedWinsor"),col=c(1,2),lty=c(1,2)
    ,lwd=2)
```



Listing 7: Gráfica Sesgo (Gamma).

```
#Gamma
plot(n,sesgo[(dis=="G")&(est=="MedArit")],ylim=c(-0.2,0.05),type="l"
    ,col=1,lty=1,lwd=2,
main="Sesgo: Dist. Gamma",xlab="Tamano de Muestra",ylab="Sesgo")
lines(n,sesgo[(dis=="G")&(est=="MedWinsor")],col=2,lty=2,lwd=2)
legend("topright",c("MedMuestral","MedWinsor"),col=c(1,2),lty=c(1,2)
    ,lwd=2,inset=0)
```



Listing 8: Tabla ECM.

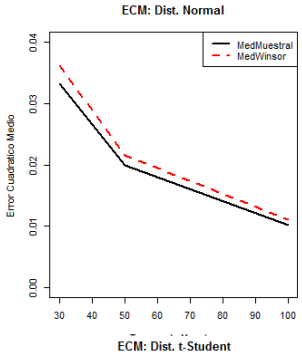
```
tab_ecm = rbind(matrix(ecm[dis=="N"],3,2,byrow=T),matrix(ecm[dis=="t"],3,2,byrow=T),
    matrix(ecm[dis=="G"],3,2,byrow=T))
tab_ecm = cbind(rep(n,3),tab_ecm)
colnames(tab_ecm) = c("T. Muestra","MedArit","MedWinsor")
rownames(tab_ecm) = rep(c("N","t","G"),each=3)
round(tab_ecm,4)
```

ECM

Escenario	Distribución	n	Media muestral	Media winsorizada
Datos simétricos sin valores *outlier*	Normal	30	0.0332	0.0362
		50	0.0199	0.0216
		100	0.0102	0.0111
Datos simétricos con valores *outlier*	t-student	30	0.0331	0.0251
		50	0.0198	0.0148
		100	0.0098	0.0074
Datos asimétricos	Gamma	30	0.0331	0.0564
		50	0.0199	0.0455
		100	0.0103	0.0392

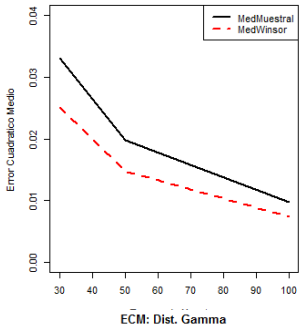
Listing 9: Gráfica Sesgo (Normal).

```
par(mfrow=c(1,3))
#Normal
plot(n,ecm[(dis=="N")&(est=="MedArit")],ylim=c(0,0.05),type="l",col
    =1,lty=1,lwd=2,
main="ECM: Dist. Normal",xlab="Tamano de Muestra",ylab="Error
    Cuadratico Medio")
lines(n,ecm[(dis=="N")&(est=="MedWinsor")],col=2,lty=2,lwd=2)
legend("topright",c("MedMuestral","MedWinsor"),col=c(1,2),lty=c(1,2)
    ,lwd=2)
```



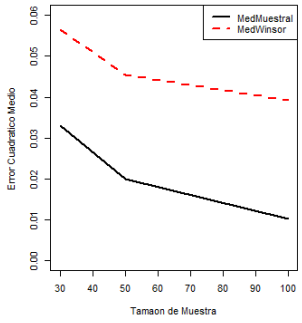
Listing 10: Gráfica Sesgo (t-Student).

```
#t-Student
plot(n,ecm[(dis=="t")&(est=="MedArit")],ylim=c(0,0.1),type="l",col
    =1,lty=1,lwd=2,
main="ECM: Dist. t-Student",xlab="Tamano de Muestra",ylab="Error
    Cuadratico Medio")
lines(n,ecm[(dis=="t")&(est=="MedWinsor")],col=2,lty=2,lwd=2)
legend("topright",c("MedMuestral","MedWinsor"),col=c(1,2),lty=c(1,2)
    ,lwd=2)
```



Listing 11: Gráfica Sesgo (Gamma).

```
#Gamma
plot(n,ecm[(dis=="G")&(est=="MedArit")],ylim=c(0,0.1),type="l",col
    =1,lty=1,lwd=2,
main="ECM: Dist. Gamma",xlab="Tamaon de Muestra",ylab="Error
    Cuadratico Medio")
lines(n,ecm[(dis=="G")&(est=="MedWinsor")],col=2,lty=2,lwd=2)
legend("topright",c("MedMuestral","MedWinsor"),col=c(1,2),lty=c(1,2)
    ,lwd=2)
```



Interpretación:
En relación al **sesgo**, para el caso de la media muestral, notamos que este disminuye a medida que aumenta el tamaño muestral en cualquiera de los tres escenarios.
En el primer escenario para datos simétricos sin valores atípicos, se observa que los sesgos son pequeños y en la mayoría de los casos la media winzorizada tiene un menor sesgo.
En el segundo escenario para datos simétricos con valores atípicos, se observa que la media muestral tiene un menor sesgo para tamaños de muestra n=30 y 50, a diferencia del tamaño de muestra $n = 100$.
En el tercer escenario, para datos asimétricos, se observa que los sesgos para la media muestral es pequeño con diferencia significativa para el caso de la media winzorizada que parece manterse constante a lo lejos del sesgo de la media muestral.

En el caso del **error cuadrático medio**, notamos que a medida que el tamaño de la muestra aumenta, el ECM disminuye en los tres escenarios vistos.
En primer escenario, los datos simétricos sin valores atípicos, observamos que el ECM es menor siempre en el caso de la media muestral. Lo que confirma que la media muestral es el estimador con menos variancia en el caso de datos normales.
En el segundo escenario, datos simétricos con valores atípicos, vemos que la media winzorizada es la que presenta menor valor, y eso es debido a que elimina el efecto de los atípicos, lo que la media muestral no puede realizar.
En el tercer escenario, los datos asimétricos, vemos que el ECM es menor en la media muestral.
□

Ejercicio 2

Estudio de simulación sobre la influencia de valores atípicos en la estimación del modelo de regresión lineal.
Sea un modelo de regresión lineal simple

$$\begin{aligned} y_i &\overset{ind}{\sim} N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

donde β_0, $\beta_1 \in \mathbb{R}$, $\sigma^2 > 0$ y x_i, $i = 1, \ldots, n$ son consideradas constantes conocidas. Para el estudio de simulación considere $\beta_0 = 1$, $\beta_1 = 2$, $\sigma = 0.2$, $n = 50$ y los valores de x_i deben ser simulados solamente una vez de una distribución *Uniforme*$(0, 1)$. Luego de simular un conjunto de datos se introducirán dos valores outlier mediante el siguiente procedimiento

- Reemplazar los dos primeros valores simulados por

$$y_i^* = y_i + 10 \times \sigma, \ i = 1, 2.$$

Estudie mediante simulación el desempeño del usual estimador de mínimos cuadrados para β_0 y β_1 en el escenario planteado. Compare sus resultados con los que se obtendrían teóricamente en el caso que no existieran valores atípicos. Interprete sus resultados.

Solución. Sabemos que el estimador de Monte Carlo del Error Cuadrático Medio de β_0 y β_1 es

$$\widehat{ECM}(T_{ij}) = \frac{1}{M}\sum_{j=1}^{M}(T_{ij} - \beta_i)^2 = \widehat{V}(T_{ij}) - \widehat{sesgo}(T_{ij})^2, \ \ i = 1, 2$$

donde T_{ij} son estimadores de prueba de β_i, además

$$\begin{aligned} \widehat{V}(T_{ij}) &= \frac{1}{M}\sum_{j=1}^{M}(T_{ij} - \bar{T}_{ij})^2 \\ \widehat{sesgo}(T_{ij}) &= \frac{1}{M}\sum_{j=1}^{M} T_{ij} - \beta_i \end{aligned}$$

Primero definimos ciertas condiciones que serán implementadas en el estudio de la simulación.

Listing 12: Definiciones iniciales.

```
# Numero de simulaciones
M=1000

# Tamano de muestra
n=50

# Parametros
beta0=1
beta1=2
sigma=0.2

# Matriz de resultados
T<-matrix(0,M,2*2)
colnames(T)<-c("beta0.Sout","beta1.Sout","beta0.out","beta1.out")
```

Se procederá a realizar la simulación de datos mediante las siguientes condiciones:

- **Escenarios:**
  (a) sin valores *outlier*: $y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$ donde $x_i \sim U(0,1)$.
  (b) con valores *outlier*: Generamos los siguientes valores atípicos: $y_i^* = y_i + 10 \times \sigma$, $i = 1, 2$.
- **Tamaño de muestra:** $n = 50$.

Listing 13: Estudio de simulación.

```
# Generando valores x (una simulacion)
x=runif(n)

# Generando valores y (1000 simulaciones)
for(j in 1:M){
        mu=beta0+beta1*x
        y<-rnorm(n)*sigma+mu

        # Sin valores outlier
        regre=lm(y~x)

        # Con valores outlier
        y[1:2]=y[1:2]+10*sigma
        regre2=lm(y~x)
        T[j,]=c(regre$coefficients[1],regre$coefficients[2],
                        regre2$coefficients[1],regre2$coefficients[2])
}
```

Finalmente se harán las evaluaciones correspondientes del desempeño del Error Cuadrático Medio para β_0 y β_1.

Listing 14: Análisis de resultados.

```
sesgo=numeric(4)
for(i in 1:2){
        sesgo[2*i-1]=colMeans(T)[2*i-1]-beta0
        sesgo[2*i]=colMeans(T)[2*i]-beta1
}
variancia=diag(var(T))

# Estimador de Monte Carlo del ECM
ecm=variancia+sesgo**2
ecm
```

Resultado:

Error Medio Cuadrático

Escenario	β_0	β_1
sin valores *outlier*	0.003694595	0.010478327
con valores *outlier*	0.060819815	0.098765976

Interpretación:
El análisis hecho es respecto al ECM de los coeficientes de regresión de la ecuación de regresión $y = \beta_0 + \beta_1 x$. Notamos que tanto para el coeficiente β_0 como para β_1, el ECM es menor cuando no existen valores outlier que cuando si existen comprobando que el efecto de valores outlier ocasiona un ECM mayor en ambos coeficientes.
Concluimos un mejor rendimiento en el escenario "sin valores *outlier*", lo cual es teóricamente correcto. □

Ejercicio 3

Una variable aleatoria X, definida en toda la recta, tiene distribución mixtura de dos normales, si su función de densidad es dada por la siguiente expresión:

$$f(x) \quad = \quad p\phi(x|\mu_1, \sigma_1^2) + (1-p)\phi(x|\mu_2, \sigma_2^2)$$

donde $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1^2, \sigma_2^2 > 0$, $p \in (0,1)$ y $\phi(\cdot|a, b^2)$ es la densidad de una distribución normal con media a y varianza b^2. Considere los datos de tiempo de espera entre erupciones del géiser Old Faithful en el Yellowstone National Park, que se encuentran en la variable waiting del conjunto de datos faithful del R, para las siguientes preguntas

a) Estime por máxima verosimilitud los parámetros de esta distribución.

Solución. Determinamos primero su función de verosimilitud

$$\begin{aligned}
\mathcal{L}(p, \mu_1, \sigma_1, \mu_2, \sigma_2) &= \prod_{i=1}^{n} f(y_i \mid \mu_1, \sigma_1, \mu_2, \sigma_2) \\
&= \prod_{i=1}^{n} \left(p\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_1^2}} + (1-p)\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_1^2}} \right) \\
&= \prod_{i=1}^{n} \left(\frac{p\sigma_2 e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_1^2}} + (1-p)\sigma_1 e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_2^2}}}{\sqrt{2\pi}\sigma_1\sigma_2} \right)
\end{aligned}$$

La función log-verosimilitud vendría dada por:

$$\ell(p, \mu_1, \sigma_1, \mu_2, \sigma_2) \quad = \quad \sum_{i=1}^{n} \log\left(p\sigma_2 e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_1^2}} + (1-p)\sigma_1 e^{-\frac{1}{2}\frac{(y_i-\mu_1)^2}{\sigma_2^2}} \right) - \log(\sqrt{2\pi}\sigma_1\sigma_2)$$

Listing 15: EMV.

```
y=faithful$waiting
p=runif(1)
u1=54
u2=80
sigma1=5.61
sigma2=6
parametros=c(p,u1,u2,sigma1,sigma2)
f1=function(parametros,y){
        a1=dnorm(y,mean = parametros[2],sd=parametros[4])
        a2=dnorm(y,mean = parametros[3],sd=parametros[5])
        -sum(log(parametros[1]*a1+(1-parametros[1])*a2))
}
r=nlminb(parametros,f1,y=faithful$waiting,lower = 0.1,upper =100)
g=r$par
g
```

Parámetros estimados

p	μ_1	μ_2	σ_1	σ_2
0.3608861	54.6148538	80.0910705	5.8712177	5.8677341

□

b) Calcule el AIC para este modelo y compárelo con el obtenido por un modelo normal.

Listing 16: AIC (Mixtura de dos normales).

```
mu=mean(y)
sigma=sd(y)
parametros2=c(mu,sigma)
f2=function(parametros2,y){
        a1=dnorm(y,mean = mu,sd=sigma)
        -sum(log(a1))
}
r2=nlminb(parametros2,f2,y=faithful$waiting,lower = 43,upper =100)
g2=r2$par
AIC2 = 2*f2(g,y)+4
AIC2
```

Listing 17: AIC (Normal).

```
AIC = 2*f1(g,y)+10
AIC
```

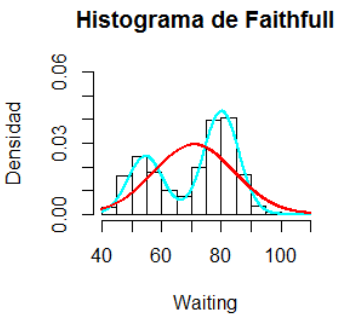
Distribución	AIC
Mixtura de dos Normales	2078.003
Normal	2194.579

Dado que el AIC de la v.a con dos dos normales es menor que el AIC de una normal, entonces el primero es un mejor modelo.

c) Realice un histograma y compare ambos modelos. Comente sus resultados.

Listing 18: AIC (Normal).

```
S = seq(from = 40, to = 110, by = 0.1)
d1 = dnorm(S, mean = g[2], sd = g[4])
d2 = dnorm(S, mean = g[3], sd = g[5])
f = g[1] * d1 + (1 - g[1]) * d2
hist(y, probability = TRUE, xlim = range(S), ylim = c
    (0, 0.06), xlab = "Waiting", ylab = "Densidad",
    main = "Histograma de Faithfull")
lines(S,f, lwd = 2,col=5)
lines(S, dnorm(S, mean = mean(y), sd = sd(y)),lwd = 2,
    col=2)
```



Según el gráfico se ratifica lo que indicaba el AIC, que la función que es una mixtura de dos normales ajusta mejor los datos y es un mejor modelo que la función normal para explicar el comportamiento de este tipo de variables que tienen dicho comportamiento, como por ejemplo la variable waiting que mide el tiempo de espera entre erupciones del geiser Old Faithful en el Yellowstone National Park.
La función normal podría quizás podría explicar en cierta medida el comportamiento de las observaciones que se encuentra en ambas colas (derecha e izquierda), pero no en la parte central, donde el error de estimación sería grande.

Ejercicio 4

Modelo de regresión lineal t de Student.
El modelo de regresión lineal t de Student es dado por

$$\begin{aligned} y_i &\overset{ind}{\sim} t(\mu_i, \sigma^2, \nu) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

donde β_0, $\beta_1 \in \mathbb{R}$, $\sigma^2 > 0$ y x_i, $i = 1, \ldots, n$ son consideradas constantes conocidas.

a) Estime por máxima verosimilitud el modelo de regresión lineal t de Student, considerando que el parámetro de grados de libertad es fijo e igual a 3 ($\nu = 3$), para las variables x_3 y y_3 del conjunto de datos de anscombe de R. Compare sus resultados con el usual modelo de regresión lineal en un gráfico de dispersión. Interprete sus resultados.

Solución. Para una variable $Y \sim t(\mu, \sigma^2, \nu)$, su función de densidad está definida como:

$$f(Y = y \mid \mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

Entonces la función de verosimilitud, siendo ν fijo, para este caso viene dada por:

$$\begin{aligned} \mathcal{L}(\beta_0, \beta_1, \sigma^2) &= f(x_1, \ldots, x_n, y_1, \ldots, y_n \mid \beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left(1 + \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \end{aligned}$$

Para obtener la estimación de máxima verosimilitud de cada parámetro basta maximizar la función log-verosimilitud:

$$\ell(\beta_0, \beta_1, \sigma) = -\left(\frac{\nu+1}{2}\right)^n \sum_{i=1}^{n}\left(1 + \frac{(y_i - \beta_0 - \beta_1 x_i)2}{\nu\sigma^2}\right) - n\log(\sigma) + cte$$

Procedemos a maximizar $\ell(\beta_0, \beta_1, \sigma)$ usando *optimización directa*:

Listing 19: EMV (Método de Optimización directa).

```
# Tomando datos x3, y3 de anscombe
x=anscombe$x3; y=anscombe$y3
nu=3;n=length(x)

# Funcion logVerosimilitud
llike =function(params){
        b0=params[1]; b1=params[2]; sigma=params[3]
        ll=n*log(sigma)+0.5*(nu+1)*sum(log(1+((y-b0-b1*x)**2)/(nu*sigma**2)))
}

# Tomamos estimadores iniciales de un modelo regresion lineal
mln=lm(y~x)
inicial=c(mln$coefficients[1],mln$coefficients[2],sd(mln$residuals))

# Optimizacion directa
mlt=nlminb(inicial,llike)

# Comparacion
inicial; mlt$par
```

Resultado:

Estimadores

Modelo	β_0	β_1	σ
Regresión Lineal	3.0024545	0.4997273	1.1728679
t-student con 3 grados de libertad	4.006308050	0.345313169	0.003071092

Mostramos el gráfico de dispersión de los datos y comparamos los modelos resultantes:

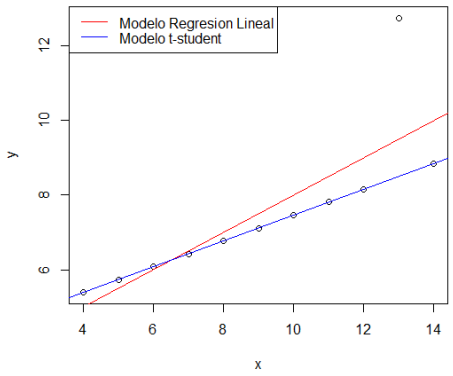
Listing 20: Gráfico de dispersión y ajuste por modelos.

```
# Grafica de dispersion
plot(x,y)

# Modelo de Regresion Lineal
abline(inicial[1],inicial[2],col="red")

# Modelo t-student con 3 grados de libertad
abline(mlt$par[1],mlt$par[2],col="blue")

legend("topleft",legend=c("Modelo Regresion
    Lineal","Modelo t-student"),lty=1,col=c("
    red","blue"))
```



La pendiente estimada bajo el supuesto de normalidad, mediante el modelo de regresión lineal, es influenciada por el valor atípico. Por su parte, el modelo t-student logra capturar perfectamente la tendencia lineal mostrada por la mayoría de observaciones. □

b) Repite el estudio de simulación de la pregunta 2, estimando ahora los coeficientes de regresión utilizando el modelo de regresión t-Student y compare sus resultados con los obtenidos en la pregunta 2.

Solución. Nos situaremos en el Escenario con valores *outlier*, en donde se generarán valores $y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$ donde $x_i \sim U(0,1)$ y los valores atípicos $y_i^* = y_i + 10 \times \sigma$, $i = 1, 2$. Simularemos β_0, β_1, con el procedimiento siguiente.

- Usando el método de regresión lineal. (Desarrollada en el problema 2).
- Usando el método de regresión t-student con 3 grados de libertad.

Listing 21: Definiciones Adicionales.

```
# Numero de simulaciones
M=1000

# Tamano de muestra
n=50

# Parametros
beta0=1; beta1=2; sigma=0.2

# Matriz de resultados
T<-matrix(0,M,2)
colnames(T)<-c("beta0.t","beta1.t")
```

Listing 22: Estudio de simulación.

```
# Generando valores x (una simulacion)
x=runif(n)

# Generando valores y (1000 simulaciones)
for(j in 1:M){
        mu=beta0+beta1*x
        y<-rnorm(n)*sigma+mu
        y[1:2]=y[1:2]+10*sigma

        # EMV de beta0 y beta1 (Optimizacion directa para t-student)
        mln=lm(y~x)
        inicial=c(mln$coefficients[1],mln$coefficients[2])
        mlt=nlminb(inicial,llike)

        T[j,]=c(mlt$par[1],mlt$par[2])
}
```

Listing 23: Análisis de resultados.

```
sesgo=numeric(2)
sesgo[1]=colMeans(T)[1]-beta0
sesgo[2]=colMeans(T)[2]-beta1
variancia=diag(var(T))

# Estimador de Monte Carlo del ECM
ecm=variancia+sesgo**2
ecm
```

Resultado: Tabla comparativa de ECM, en Escenario con valores *outlier* respecto a los métodos de regresión lineal y regresión t-student.

Error Medio Cuadrático

Modelos	β_0	β_1
Modelo de regresión lineal	0.060819815	0.098765976
Modelo t-student	0.01668002	0.01541430

Interpretación:
Nos situamos en un escenario con valores outlier, y observamos que el ECM obtenido tanto para β_0 como para β_1 mediante el modelo de regresión t-student es menor en ambos casos con respecto a lo obtenido usando el modelo de regresión lineal vista en el Ejercicio 2.
Esto nos indica un mejor rendimiento por parte del Modelo de regresión t-student.

□