Estadística y modelación de sistemas socioecológicos en R

LANCIS | Laboratorio Nacional de Ciencias de la Sostenibilidad

Dra. Yosune Miquelajauregui Graf

Plan del día

1. Criterio de información de Akaike (AIC): selección de modelos e inferencia multimodelo

2. Máxima verosimilitud

3. Ejercicios

4. Disusión sobre análisis de datos de tesis y seminario

AIC: selección de modelos

Criterio de información de Akaike: Es una medida de la calidad relativa de un modelo estadístico, para un conjunto de datos.

$$\mathbf{AIC = -2 \log (L) + 2K}$$

donde L es una medida del ajuste del modelo dada por el valor máximo de la función de verosimilitud y K es el número de parámetros a estimar.

Máxima verosimilitud

- La función de verosimilitud es una herramienta desarrollada por Sir Ronald Fisher (1890-1962) utilizada en la resolución de múltiples problemas.

- Esta herramienta se basa en un proceso iterativo, es decir, se encuentran los valores de los parámetros que maximicen la verosimilitud de la función.

- En otras palabras, se encuentran los valores más probables de los parámetros dados los datos $L(\theta|x)$, donde θ representa uno o más parámetros y x representa una variable o un juego de datos.

- $L(\theta|x)$ indica entonces, la probabilidad de observar el valor de θ para cierto juego de datos.

Máxima verosimilitud

- La función de verosimilitud corresponde también a la función de densidad de probabilidad de una variable aleatoria.

- La función de densidad describe la probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor.

$$\boxed{L(\theta \mid x)} = \boxed{P(x \mid \theta)}$$

Función de máxima verosimilitud

Función de densidad de una variable aleatoria

Máxima verosimilitud: Distribución discreta

- Imaginemos un experimento con una moneda. Queremos saber la probabilidad de obtener cara. Lanzamos 20 veces la moneda y obtenemos lo siguiente:

lanzamiento [1]
"cruz" "cara" "cara" "cara" "cara" "cara" "cara" "cruz" "cruz" "cara” "cara" "cara"
"cara" "cara" "cara” "cara" "cruz" "cara" "cruz" "cara"

table(lanzamiento)
cara cruz
 15   5

Si alguien preguntara sobre la probabilidad de obtener “cara”, podríamos responder que ésta es ciertamente, mayor que la probabilidad de obtener “cruz”.

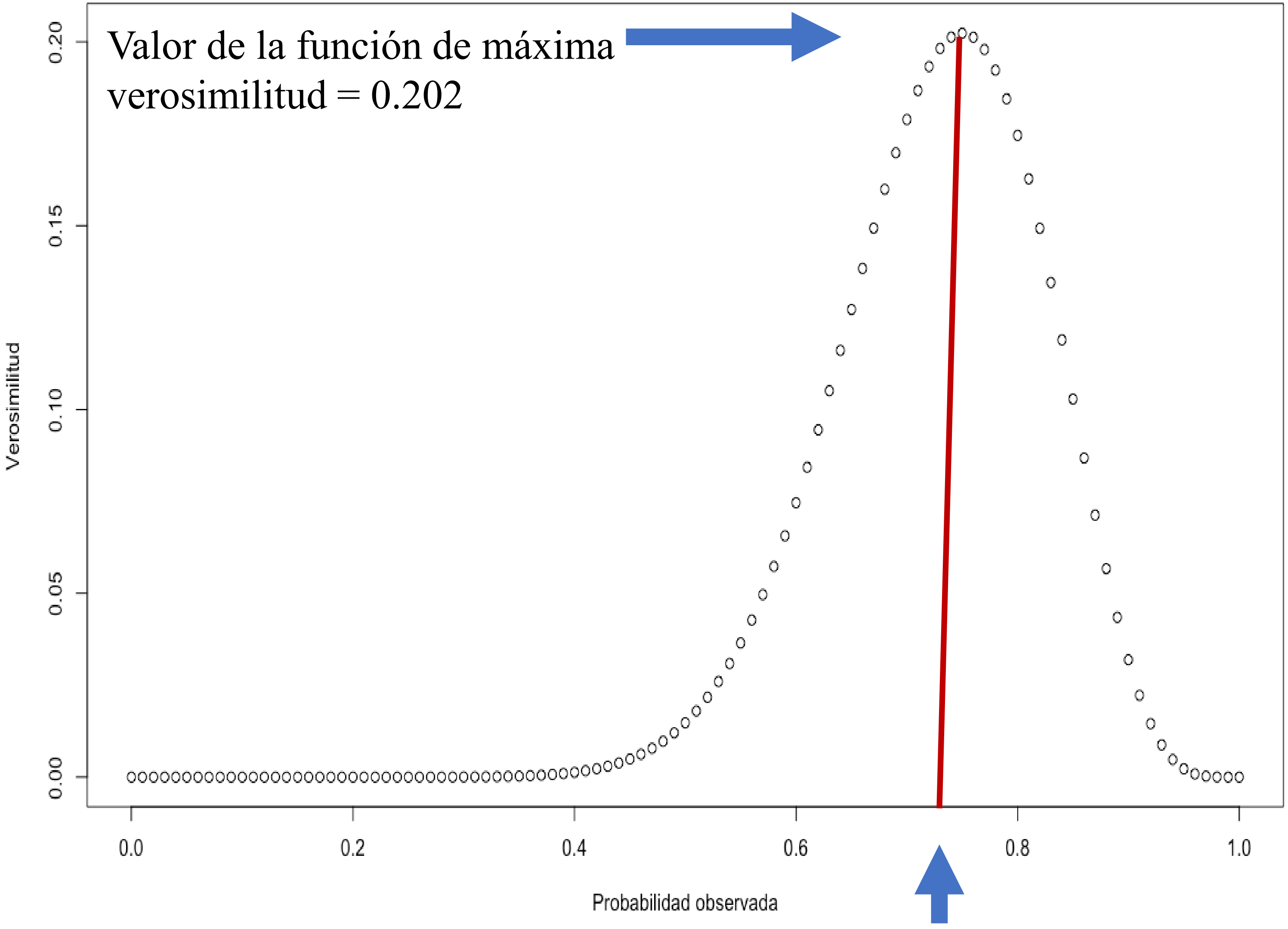
Máxima verosimilitud: Distribución discreta

- Tenemos un experimento binomial (éxito vs fracaso).

- Podemos utilizar la función de densidad para la distribución binomial: distribución discreta que cuenta el número de éxitos x (obtener "cara") en una secuencia de n ensayos (20), con una probabilidad p de obtener 1 éxito. Sabemos que p puede variar entre 0 y 1.

dbinom (x= , size= , prob=)

Máxima verosimilitud: Distribución discreta



Estimado de p en el punto máximo (MLE) = 0.75

```
theta <- seq (from=0, to =1, by=0.01)
Vero <- dbinom(x=15, size=20, prob=theta)

which(Vero==max(Vero))
76
theta[76]
0.75
```

- Es decir, si observamos x=15 "caras" sobre n=20 lanzamientos, tenemos una probabilidad de 0.75 de obtener "cara" con esta moneda.

- 0.75 es el valor más probable según la función de máxima verosimilitud dados los datos.

Máxima verosimilitud: Distribución continua

- Queremos encontrar el valor más probable de la media poblacional θ dado que una muestra de 6 observaciones proviene de una distribución normal y una desviación estándar de 3:

x <-12 3 5 8 6 4

- mean(x)=6.3

- Cada observación contiene información para estimar θ $(la\ media)$

- Propiedad de verosimilitud: Es posible combinar los L(θ) de cada juego de datos independientes y multiplicarlos

$$L(\theta)=L1(\theta)+ L2(\theta)+ L3(\theta)+ L4(\theta)+ L5(\theta)+ L6(\theta)$$

Máxima verosimilitud: Distribución noraml

- Utilizando la misma estrategia, podemos construir los logL (θ)

```
x <- c(12,3,5,8,6,4)
theta2 <- seq(from=0, to=14, by=0.01)
LL1<-log(dnorm(x=12,mean=theta2,sd=3))
LL2<-log(dnorm(x=3,mean=theta2,sd=3))
LL3<-log(dnorm(x=5,mean=theta2,sd=3))
LL4<-log(dnorm(x=8,mean=theta2,sd=3))
LL5<-log(dnorm(x=6,mean=theta2,sd=3))
LL6<-log(dnorm(x=4,mean=theta2,sd=3))
```

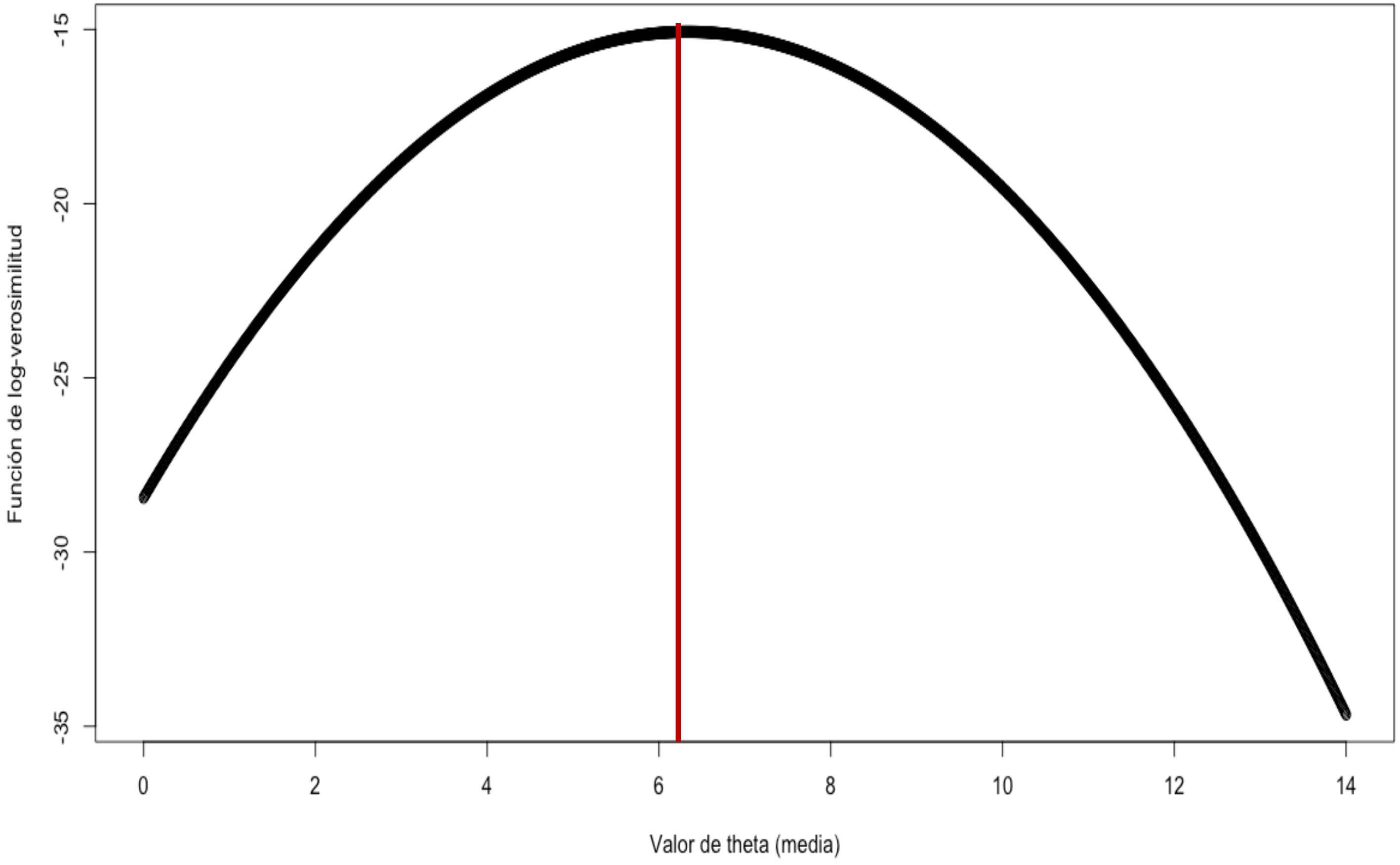
Necesitamos datos independientes!!

Máxima verosimilitud: Distribución normal

log_LL <-
LL1+LL2+LL3+LL4+LL5+LL6
max(log_LL)
theta2[which(log_LL==max(log_LL))]

6.33

La media de la muestra corresponde a $\boldsymbol{\theta}$ (media) estimada por el método de máxima verosimilitud



Máxima verosimilitud

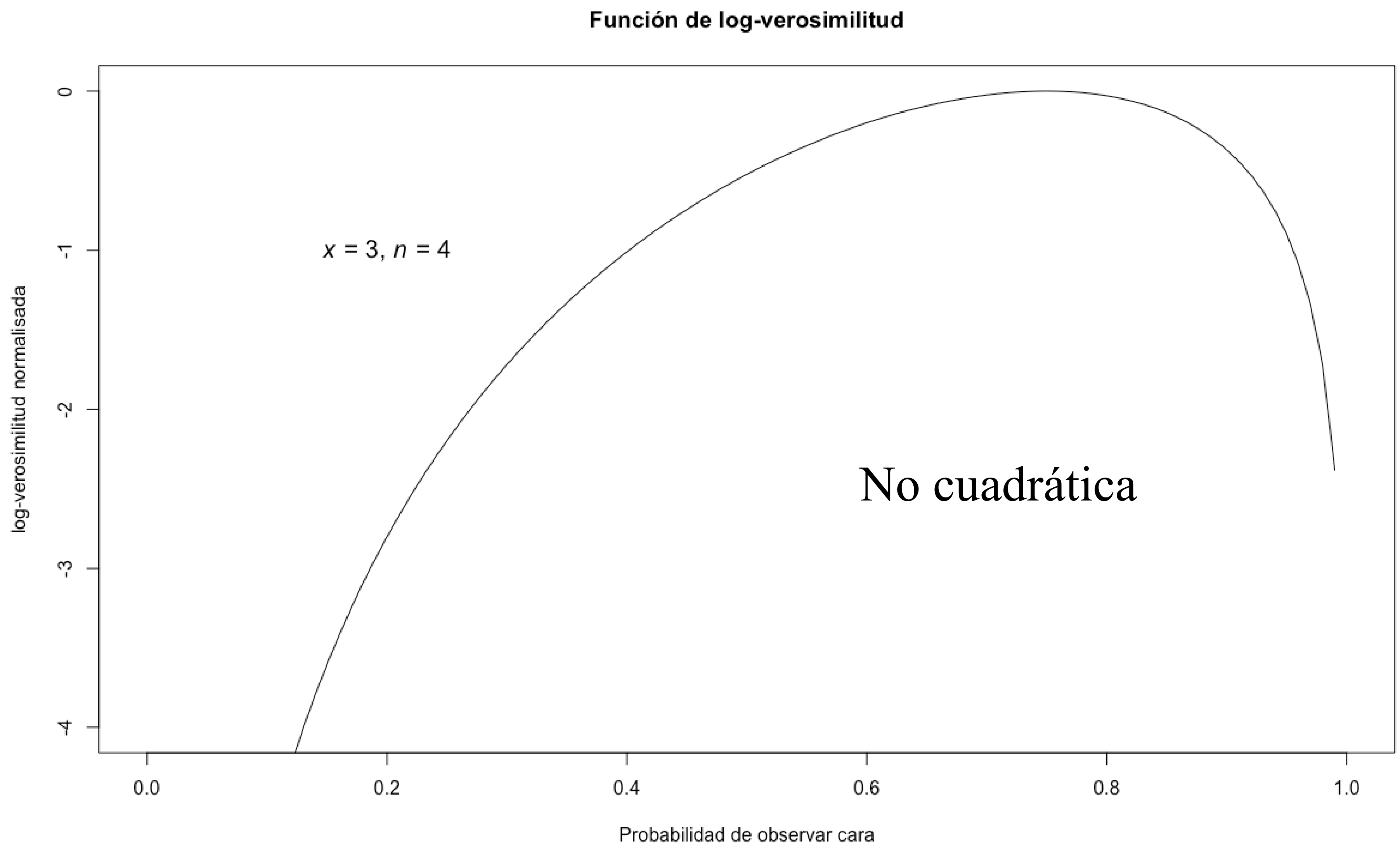
- La forma de la curva de la función de verosimilitud es importante, particularmente la de log L (logL(θ)).

- Podemos graficar la curva de la función logL para verificar si es cuadrática: principio- sustraer el máximo logL(θ) de todos los logL(θ) y verificar la forma de la curva.

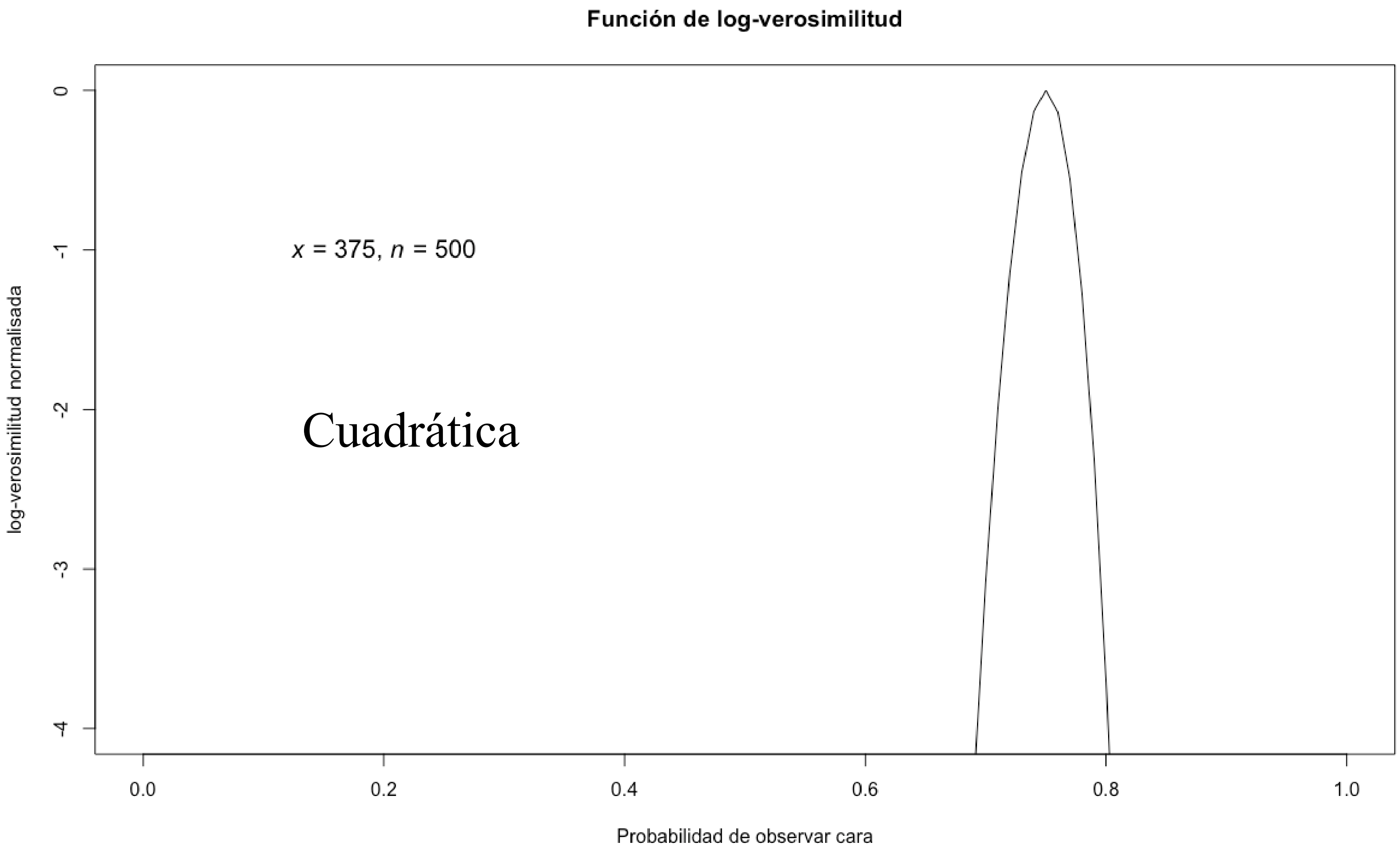
- Una curva cuadrática de logL(θ) corresponde a una aproximación normal de $\hat{\theta}$, es decir, implica que $\hat{\theta}$ tiene una distribución normal.

- Si la forma no es cuadrática, los errores estándar (medida de la precisión del estimado) no son apropiados.

Máxima verosimilitud



Máxima verosimilitud



AIC: selección de modelos

- AIC es una herramienta interesante cuando se tiene una serie de modelos pertinentes que hemos especificado *a priori*.

- El modelo con el valor de AIC más bajo es el “mejor” para los datos que alimentan a los modelos.

- AIC no es una prueba de hipótesis, no hay noción del valor de significancia. Se habla más bien del concepto de un modelo “pobre” y “mejor”.

- Especificación de modelos: Justificación biológica de modelos que incluyan o no términos.

AIC: selección de modelos

- Selección de modelos = selección de variables

¿Qué variables debo utilizar?
¿Qué modelo debo utilizar?

- Es MUY importante utilizar la misma base de datos, esto quiere decir que la variable dependniente no cambia.

- Podemos hacer variar la función de liga (link). El interior del modelo podemos utilizar transformaciones.

- Verificamos el ajuste del modelo más complejo (modelo general).

Utilización de AIC

- Para muestras pequeñas (n/k<~40) se utiliza una versión modificada de AIC en el que se agrega un factor de corrección:

$$AIC = -2 \log (L) + \underbrace{\boxed{2K + 2K(K+1)/(n-K-1)}}_{\text{Factor de corrección}}$$

Clasificación de modelos

- Podemos ordenar los modelos (entre todos los que consideremos) en relación al mejor modelo.

$$\text{Delta AICi} = \Delta i = \boxed{AICi} - \boxed{minAIC}$$

Modelo i

Mejor modelo
(con el AIC más bajo)

Clasificación de modelos

- Los valores de Delta AIC son fáciles a interpretar y proporcionan una metodología para comparar los modelos.

Delta AIC	Interpretación
<2	Modelo probable
Entre 4 y 7	Modelo poco probable
>10	Modelo improbable

Pesos de Akaike

- Podemos calcular los pesos de Akaike para cada modelo.
- Se calcula como el cociente de los delta de cada modelo entre el total de deltas.
- Los pesos de akaike se expresan en escala entre 0 y 1. La suma de los pesos es

$$\text{Pesos de Akaike} = wi= \frac{e^{\frac{-\Delta i}{2}}}{\sum_{r=1}^{R} e^{\frac{-\Delta r}{2}}}$$

Pesos de Akaike

- Interpretación del peso de Akike: la probabilidad que un modelo sea mejor teniendo en cuenta a los otros modelos candidatos.

- Por ejemplo, un peso de 0.65 para un modelo significa que tiene 65% más probabilidad de ser mejor modelo que el resto de los modelos candidatos.

- El modelo 1 es 0.65/0.22~3 veces mejor que el modelo 2 (evidencia de cociente)

Modelo	K	AIC	Delta AIC	Peso de Akaike
1	2	12.3	0	0.65
2	4	14.4	2.15	0.22

Ejemplo

- Tiempo necesario de las salamandras para recorrer 1 m en el suelo según distintas variables del micro-hábitat.

- Tiempo es una variable continua (regresión lineal)

Ejemplo

- Variables que podrían influenciar potencialmente el desplazamiento de las salamandras:

- Aire: temperatura del aire : CONTINUA
- Cobertura: Cobertura vegetal (<50 m vs >50 m) : BINARIA
- Largo: Largo de la salamandra (cm) : CONTINUA

- Modelos plausibles:

1) Aire
2) Cobertura
3) Largo+Cobertura
4) Aire+Cobertura+Largo
5) Largo+Cobertura+Largo*Cobertura
6) Aire+Cobertura +Largo+Cobertura*Largo

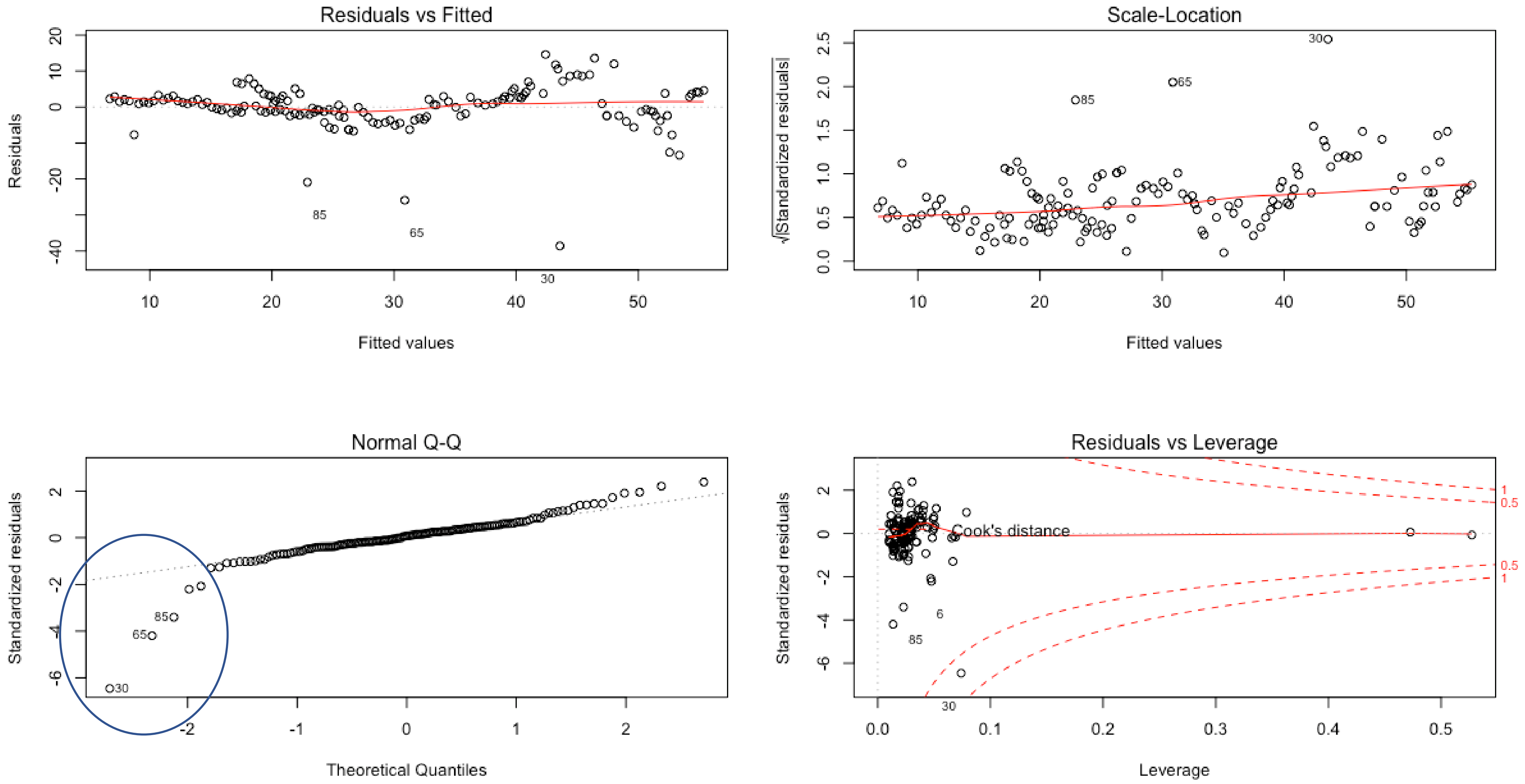
Ejemplo

1. Verificar el ajuste del modelo global

```
summary(mod1)
Call:lm(formula = Tiempo ~ Aire + Largo + Cobertura + Cobertura:Largo,     data = saltiempo)
Residuals:    Min      1Q  Median      3Q     Max -38.584  -2.385   0.525   2.890  14.591
Coefficients:
                  Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)       132.4397    5.7891       22.878    <2e-16 ***
Aire               -5.3206    0.3323      -16.010    <2e-16 ***
Largo               0.2104    0.3746        0.562     0.5752
Cobertura          13.1443    7.6162        1.726     0.0865 .
Largo:Cobertura    -0.2107    0.3746       -0.562     0.5747
 ---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
Residual standard error: 6.208 on 144 degrees of freedom
Multiple R-squared:  0.8326,
Adjusted R-squared:  0.828
F-statistic: 179.1 on 4 and 144 DF,  p-value: < 2.2e-16
```

Ejemplo

1. Verificar el ajuste del modelo global



Ejemplo

2. Obtener el valor de logL (θ)
   Función en R : logLik()

logLik(mod1)'log Lik.' -480.9214 (df=6)

El valor máximo de la función de verosimilitud es **-480.9** dados los datos

Ejemplo

3. Correrer modelos alternativos/candidatos

```
mod2<- lm(Tiempo~Aire, data=saltiempo)
mod3<- lm(Tiempo~Cobertura, data=saltiempo)
mod4<- lm(Tiempo~Largo, data=saltiempo)
mod5<- lm(Tiempo~Largo+Cobertura, data=saltiempo)
mod6<- lm(Tiempo~Largo+Cobertura+Cobertura:Largo, data=saltiempo)
mod7<- lm(Tiempo~Aire+Largo+Cobertura, data=saltiempo)
```

Ejemplo

4. Obtener AIC, Delta AIC, Pesos

```
AICc_1<--2*logLik(mod1)[1]+2*(length(coefficients(mod1))+1)+
(2*(length(coefficients(mod1))))*(length(coefficients(mod1)+1))
/(150-length(coefficients(mod1))-1)
```

```
Delta_AIC_1<-AICc_1-min(AICtotal)
```

```
Pesos AIC_1 <- exp(-Delta_AIC_1/2)/sum(exp(-Resultados$Delta_AICc/2))
```

Ejemplo

5. Generar tabla con resultados

	Modelo	AICc	DeltaAICc	AICcPeso
7	Aire+Largo+Cobertura	972.5865	0.000000	6.902973e-01
1	Largo+Aire+Cobertura+Largo:Cobertura	974.1901	1.603581	3.096158e-01
2	Aire	990.5454	17.958969	8.695523e-05
6	Largo+Cobertura+Largo:Cobertura	1124.6028	152.016295	6.747099e-34
3	Cobertura	1157.4813	184.894853	4.893606e-41
5	Largo+Cobertura	1159.5821	186.995641	1.711783e-41
4	Largo	1232.7768	260.190317	2.184876e-57

El modelo 7 es 0.69/0.39= 1.7 más probable de ser mejor modelo que el modelo 1 y ~9 veces mejor que el modelo 2

Inferencia multimodelo

- Cuando tenemos múltiples modelos, podemos basar nuestra inferencia sobre el modelo que se encuentra en la primera posición, siempre y cuando su peso >0.90

- Cuando el peso <0.90, distintos modelos son plausibles.

- En este caso, la mejor estrategia es basar la inferencia sobre el conjunto de modelos candidatos (Inferencia multimodelo).

Inferencia multimodelo

Calcular el promedio de modelos

- En lugar de fiarse solamente en las estimaciones del mejor modelo, podemos basar nuestros cálculos en el promedio ponderado de todas las estimaciones derivadas de todos los modelos.

- Robusto y preciso:

$$\hat{\hat{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

donde w_i= pesos de Akaike y $\hat{\theta}_i$= estimadores modelo i

Inferencia multimodelo

Calcular el promedio de modelos

- Imaginemos que estamos interesados en el efecto de la variable Aire sobre el tiempo de desplazamiento de las salamandras.

- Tres modelos plausibles:

1) Aire
2) Aire+Cobertura+Largo
3) Aire+Cobertura +Largo+Cobertura*Largo

Inferencia multimodelo

Recalcular los Delta AIC y los pesos de Akaike para los modelos que incluyen la variable de interés: Aire

	Modelo	AICc	DeltaAICc	AICcPeso
7	Aire+Largo+Cobertura	972.5865	0.000000	6.902973e-01
1	Largo+Aire+Cobertura+Largo:Cobertura	974.1901	1.603581	3.096158e-01
2	Aire	990.5454	17.958969	8.695523e-05

Inferencia multimodelo

Calcular el promedio de los modelos para la variable Aire

```
aire.ests <- c(coef(mod1)[2], coef(mod2)[2], coef(mod7)[2])
aire.ests     Aire        Aire        Aire
          -5.320552  -4.149320   -5.212568
```

```
mod.avg.est <- sum(ResultadosA$AICcPeso*aire.ests)
```

```
mod.avg.est[1]
```

-5.245909

Media ponderada de la variable Aire

Inferencia multimodelo

Calcular la precisión de la media ponderada

- Podemos calcular también el SE de las estimaciones, SE incondicional (Anderson 2008)

$$SE = \sqrt{\sum_{i=1}^{R} w\,i\,\widehat{var}\left(\hat{\theta}_i \middle| g_i\right) + (\hat{\theta}_i - \hat{\hat{\theta}}_i)^2}$$

donde wi son los pesos de Akaike, $\widehat{var}$ es la varianza estimada, $\hat{\theta}_i$= estimador en cuestión (aire) del modelo i, $\hat{\hat{\theta}}_i$ es la media ponderada del estimador en cuestión

Inferencia multimodelo

```
aire.se <- c(summary(mod1)$coef[2, 2], summary(mod2)$coef[2, 2],
summary(mod7)$coef[2, 2])
```

aire.se[1] 0.3323302 0.1682901 0.2706232

```
incond.se <- sum(ResultadosA$AICcPeso*sqrt((aire.se^2) + ((aire.ests -
mod.avg.est)^2)))
```

incond.se[1] **0.2937775**

Inferencia multimodelo

- Una vez que tenemos la media ponderada y el error estándar podemos construir los intervalos de confianza para evaluar el efecto de la variable Aire.

- Para el efecto de la variable Aire:

$$-5.245909 \pm 0.2937775$$

- Límite inferior 95%: **-5.82**
- Límite superior 95%: **-4.67**

Conclusión: El cero no está incluido en el intervalo, por lo tanto hay un efecto de la variable Aire

Ventajas de AIC vs Ho

- Objetivo: No se basa en el nivel de significancia (0.05)
- Riguroso y sencillo de calcular
- Permite comparar modelos
- Ideal para seleccionar modelos o variables
- Fundado sobre principios sólidos de estadística (máxima verosimilitud)
- Podemos incorporar incertidumbre (promedio ponderado)

Ejercicio

1. Importar la base de datos SO.txt que contiene información sobre la concentración de SO_2 en función de variables ambientales y el tamaño de la población.

2. Ajustar los 5 modelos lineales siguientes y realizar la selección de modelos con ayuda del criterio de información de Akaike. ¿Qué podemos concluir?

a) SO2~Temp+Industria+Viento+Lluvia+DiasHumedos
b) SO2~Temp+Poblacion+Viento+Lluvia+DiasHumedos
c) SO2~Temp+Viento+Lluvia+DiasHumedos
d) SO2~Poblacion
e) SO2~Industria

3. Efectuar la inferencia multimodelo para el efecto de Temperatura (Temp) y calcular los intervalos de confianza incondicionales (95%) asociados a este parámetro.