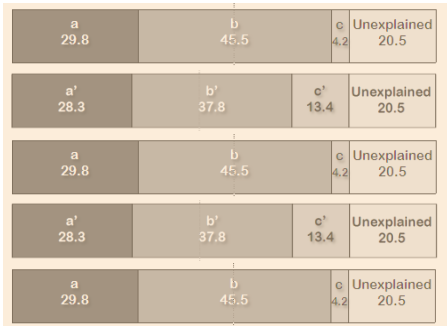
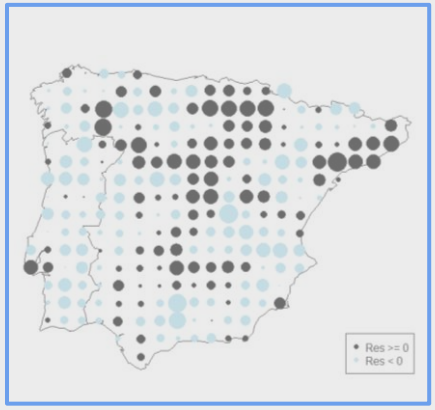




Técnicas y paquetes de R para la evaluación o diagnóstico de modelos de regresión

Dolores Ferrer Castán[1],
Jennifer Morales Barbero[1] y Ole R. Vetaas[2]

[1]Área de Ecología, Facultad de Biología, Universidad de Salamanca
[2]Departamento de Geografía, Universidad de Bergen, Noruega

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

Universitetet i Bergen

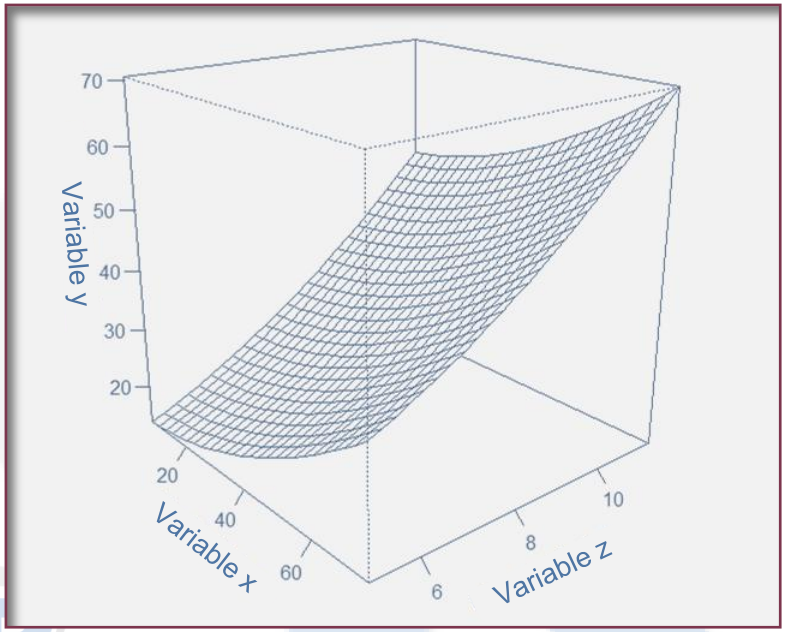
Introducción

Modelos

Representaciones abstractas y formales de la realidad que pretendemos describir, analizar y, si es posible, comprender

$$y = b_0 + b_1x + b_2x + \ldots + e$$

Modelo de regresión



$$\text{error} = y_{obs} - \hat{y}$$

El examen y la evaluación de los errores del modelo son fundamentales

De Ferrer-Castán y Vetaas (2005), redibujado

Observaciones anómalas

Comportamiento en el origen de coordenadas

Distribución de los errores

Estructura espacial de los datos

Otras técnicas de evaluación

Observaciones anómalas

¿Hay observaciones o casos anómalos?

¿Tienen una influencia excesiva los casos anómalos?

¿Olvidamos incorporar alguna tendencia en el modelo?

¿Hay observaciones o casos anómalos?

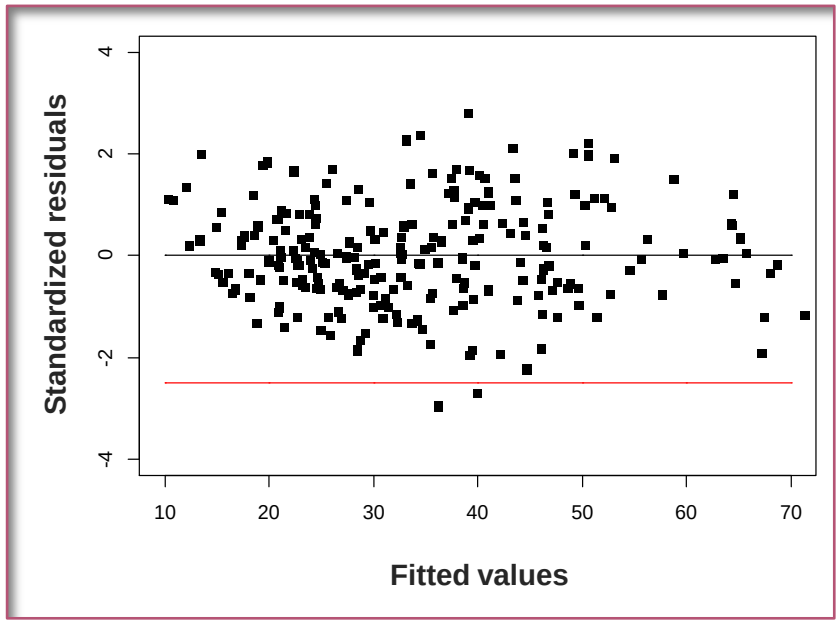
- **Observaciones anómalas** (casos raros)
- **Datos erróneos** (equivocaciones)

$$\text{error} = y_{obs} - \hat{y}$$

Residuales vs. valores teóricos

```
rawres.lm <- residuals(fit1.lm)
stdres.lm <- rstandard(fit1.lm)
fitted.lm <- fitted(fit1.lm)
```

```
plot(fitted.lm, stdres.lm,
     type="p", lty=2,
     xlim = c(10, 70),
     ylim = c(-4, 4), pch = 15,
     xlab="Fitted values",
     ylab="Standardized residuals", font.lab=2)
axis(1, pos = 0, lty=1, tck=0, labels=F)
axis(1, pos = -2.5, lty=1, col="red", tck=0, labels=F)
```



Riqueza de especies leñosas en la Península Ibérica. Modelo a 50x50km² de Vetaas y Ferrer-Castán (2008)

¿Tienen una influencia excesiva los casos anómalos?

- **Estimación de los parámetros** del modelo
- **Estadísticos de bondad de ajuste**

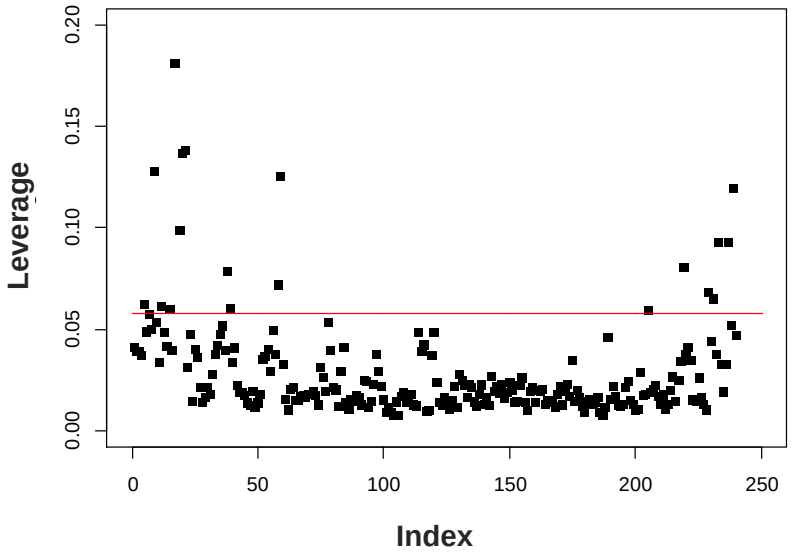
Umbral: Si >2*valor medio

[Hoaglin y Welsch (1978) y McCullagh y Nelder (1989)]

Medidas de leverage

```
lev.lm <- hatvalues(fit1.lm)

# Otra opción:
library(MASS)
lev <- hat(model.matrix(fit1.lm))
## Esta función existe sobre todo
para compatibilidad con S (versión
2); recomendada la función del
paquete base "stats"
```



Riqueza de especies leñosas en la Península Ibérica. Modelo a 50x50km² de Vetaas y Ferrer-Castán (2008)

```
plot(lev.lm, type="p", lty=1,
     pch = 15, xlab="Index", ylab="Leverage",
     font.lab=2, xlim = c(0, 250), ylim = c(0.00, 0.20))
axis(1, pos = 0.058, lty=1, col="red", tck=0, labels=F)
```

¿Tienen una influencia excesiva los casos anómalos?

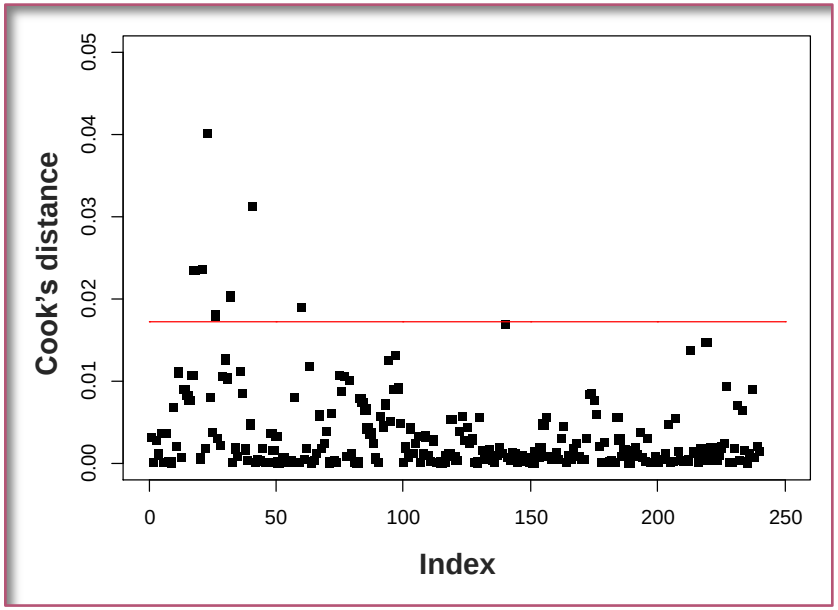
- **Estimación de los parámetros** del modelo
- **Estadísticos de bondad de ajuste**

Umbral: si >4/(N-número de parámetros modelo)

Distancias de Cook

```
cook.lm <- cooks.distance(fit1.lm)
```

```
plot(cook.lm, type="p",
     lty=2, pch=15,
     xlab="Index",
     ylab="Cook's distance",
     font.lab=2, xlim = c(0, 250),
     ylim = c(0.00, 0.05))
axis(1, pos = 0.01716738, lty=1,
     col="red", tck=0, labels=F)
```



Riqueza de especies leñosas en la Península Ibérica. Modelo a 50x50km² de Vetaas y Ferrer-Castán (2008)

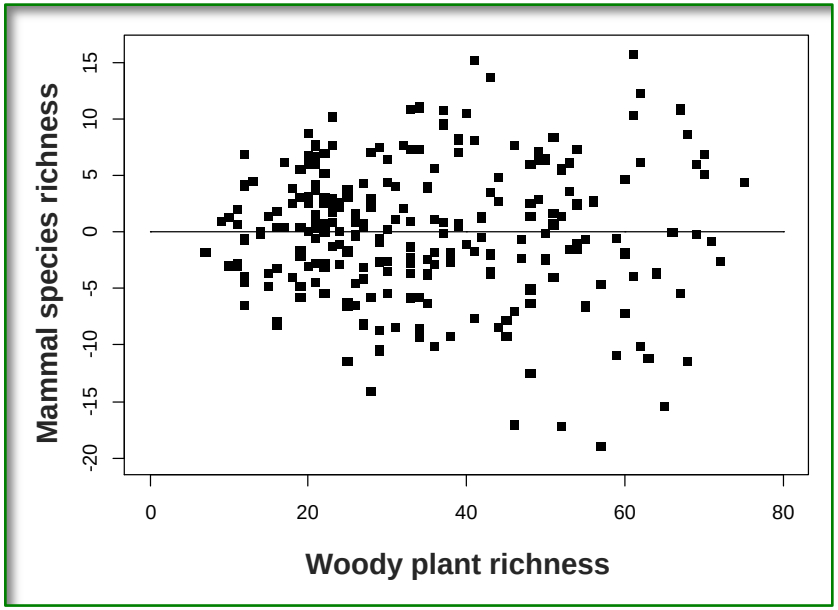
¿Olvidamos incorporar alguna tendencia en el modelo?

Residuales vs. predictores

```
rawres.lm <- residuals(fit1.lm)
a <- cbind(rawres.lm)
```

```
plot(wood, a, lty=2, pch=15,
     xlim = c(0, 80),
     ylim = c(-20, 16),
     xlab="Woody plant richness",
     ylab="Mammal species richness",
     font.lab=2)
```

```
axis(1, pos = 0, lty=1, tck=0, labels=F)
```



Comportamiento en el origen de coordenadas

¿Tiene sentido?

Otras técnicas en lugar de los modelos lineales

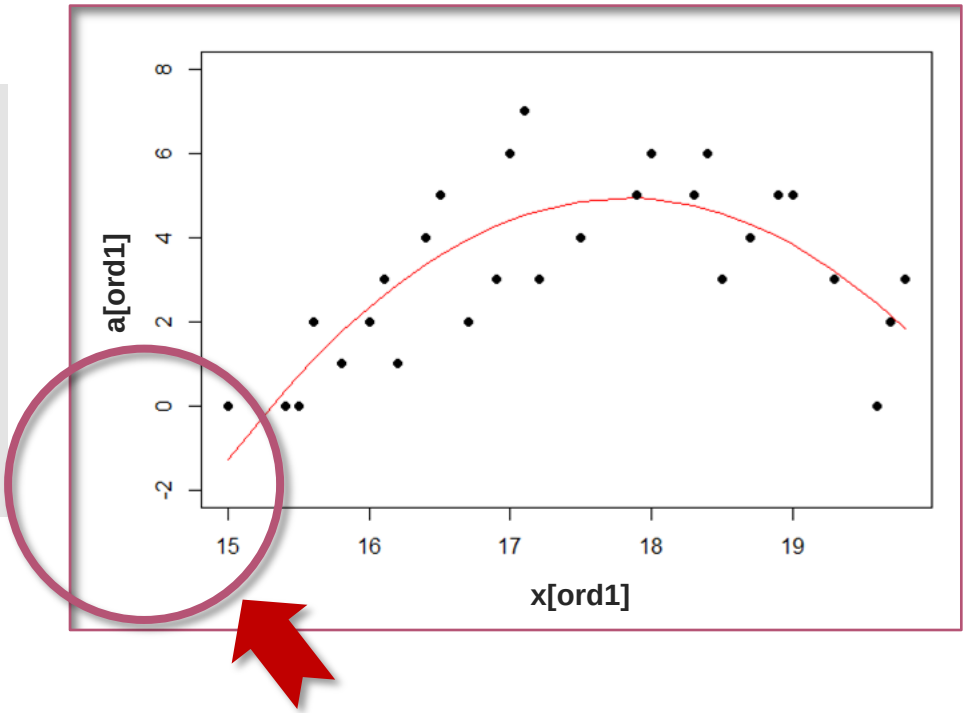
¿Tiene sentido?

Si el modelo predice **valores negativos** para la variable respuesta... ¿tiene esto sentido?

```
fit1.lm <- lm(y ~ poly(x,2))
a <- fitted(fit1.lm)

ord1 <- order(x)

plot(x[ord1],a[ord1], type="l",
     ylim=c(-2, 8), col="red")
points(x,y, pch=16)
```



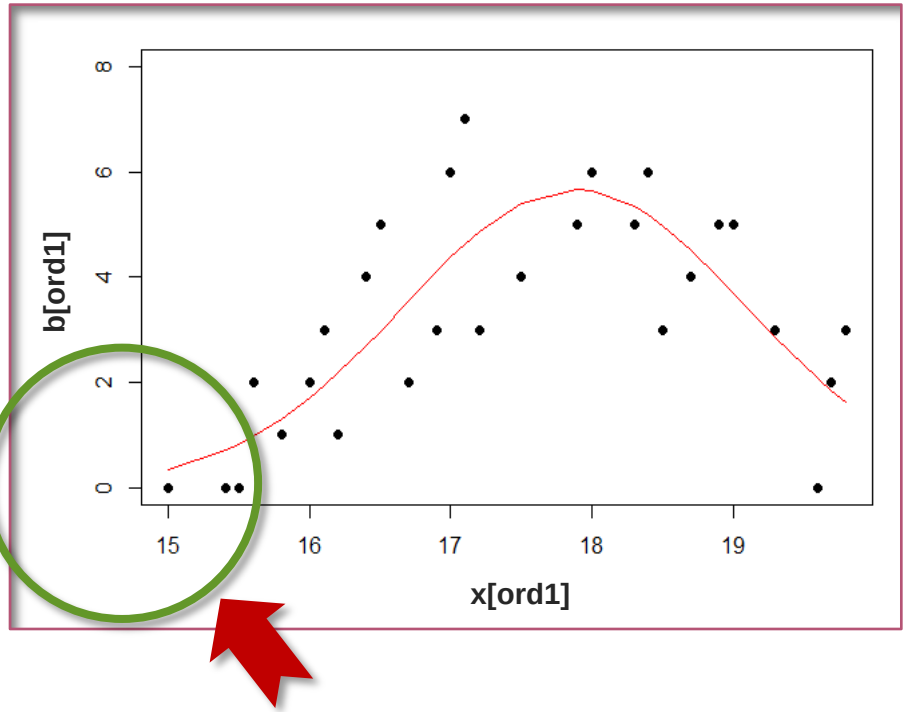
Otras técnicas en lugar de los modelos lineales

Si el modelo predice **valores negativos** para la variable respuesta... ¿tiene esto sentido?

Si no lo tiene, una opción puede ser la utilización de **modelos lineales generalizados** (**GLMs**) de la familia de Poisson (vínculo logarítmico)

```
fit1.glm <- glm(y ~ poly(x,2),
                family=poisson)
b <- fitted(fit1.glm)

ord1 <- order(x)
plot(x[ord1], b[ord1], type="l",
     ylim=c(0, 8), col="red")
points(x,y, pch=16)
```



Distribución de los errores

¿Están los errores normalmente distribuidos?

¿Muestran algún patrón geográfico?

¿Están los residuales normalmente distribuidos?

Q-Q plots

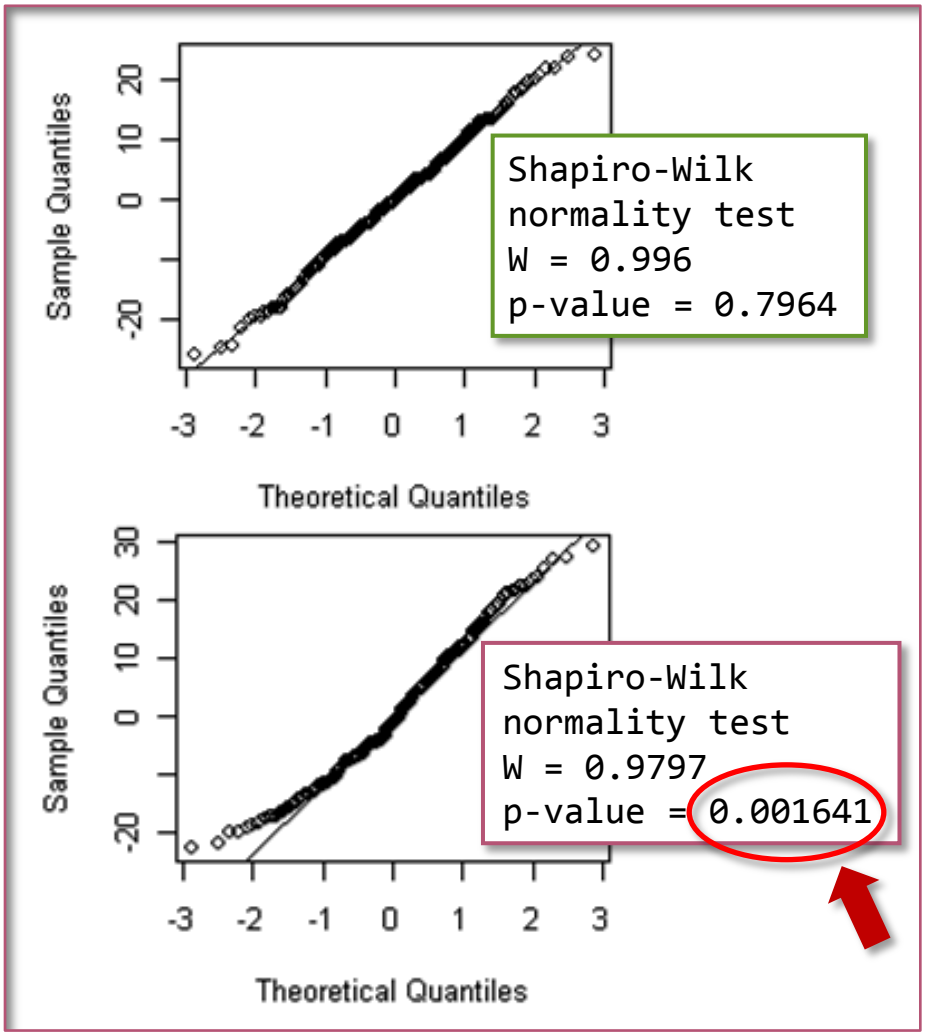
```
qqnorm(resid(fit1.lm))
qqline(resid(fit1.lm))
```

Shapiro-Wilk normality test

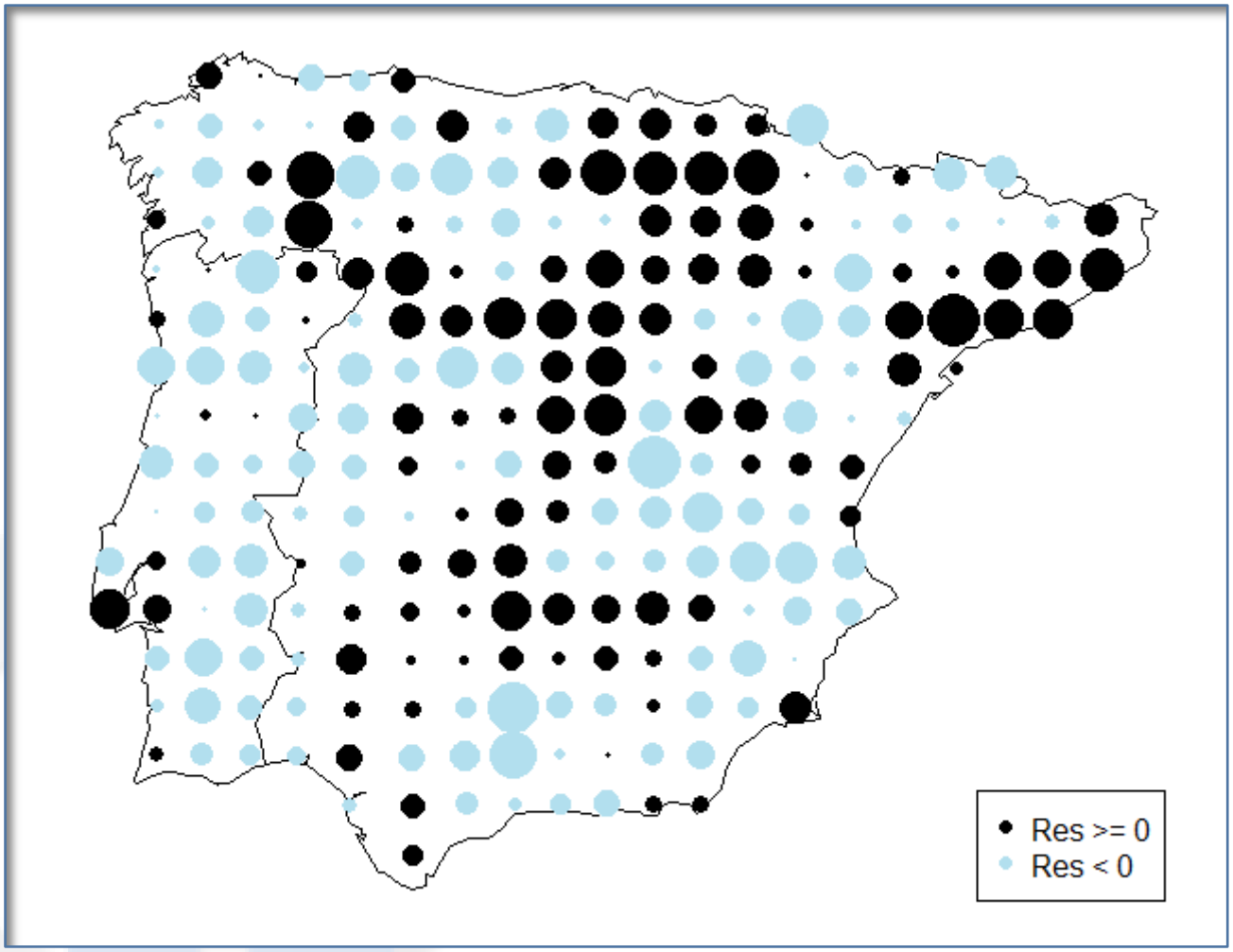
```
shapiro.test(resid(fit1.lm))
```

H_0: los residuales están normalmente distribuidos

si $P>0.05$, H_0 no puede ser rechazada -> los residuales siguen una distribución normal



¿Muestran algún patrón geográfico?



Riqueza de especies leñosas en la Península Ibérica. Residuales del modelo a 50x50km² de Vetaas y Ferrer-Castán (2008)

¿Muestran algún patrón geográfico?

```
fit1.lm <- lm(wood ~ eler+map+I(map^2)+I(map^3)+aet+map:calc)
rawres.lm <- residuals(fit1.lm)
```

```
geo.res.lm <- cbind(rawres.lm, lon, lat)
geo.res.df <-as.data.frame(geo.res.lm)
```

```
library(spdep)
coordinates(geo.res.df)<-c("lon", "lat")
```

```
library(spdep)
-> library(sp)
-> library(Matrix)
-> library(lattice)
library(maps)
library(mapdata)
```

```
library(maps)
library(mapdata)
iberia <-map("worldHires",
             regions=c("Spain", "Spain:Cabo de Palos", "Portugal", "Andorra"),
             exact=TRUE)
```

```
Nresid.lm = subset(geo.res.df, rawres.lm <0)
Presid.lm = subset(geo.res.df, rawres.lm >=0)
```

```
plot(Nresid.lm, col="lightblue2", pch=19,
     cex =sqrt(abs(Nresid.lm$rawres.lm))/1.3, add=TRUE)
plot(Presid.lm, col="black", pch=19,
     cex =sqrt(Presid.lm$rawres.lm)/1.3, add=TRUE)

legend("bottomright", c("Res >= 0", "Res < 0"), pch=19,
       col=c("black", "lightblue2"))
```

Estructura espacial de los datos

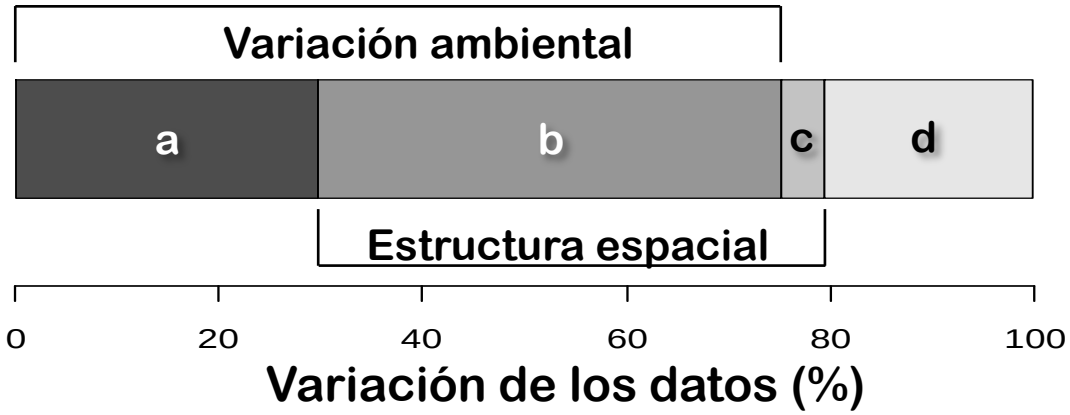
Regresiones parciales y partición de varianzas

Semivariogramas

Coeficientes de Moran I y correlogramas

Modelado espacialmente explícito

Regresiones parciales y partición de varianzas



- **a** Variación puramente ambiental
- **b** Variación ambiental espacialmente estructurada
- **c** Variación puramente espacial
- **d** Variación no explicada

Variación ambiental (a+b)

```
fit1.lm <- lm(treeFM ~ poly(eler,2)+poly(map,3)+aet+map:calc)
```

Estructura espacial (b+c)

```
fit1.lm <- lm(treeFM ~ lon+poly(lat,2))
```

Variación total (a+b+c)

```
fit1.lm <- lm(treeFM ~ lon+poly(lat,2)+poly(eler,2)+poly(map,3)+aet+map:calc)
```

a = (a+b+c)−(b+c) c = (a+b+c)-(a+b) b = (a+b)−(a) = (b+c)−(c)

```
a <- c(30, 46, 4, 20)
m <- matrix(a, nrow=4, ncol=1)
barplot(m, beside=F, horiz=T)
```

Semivariogramas

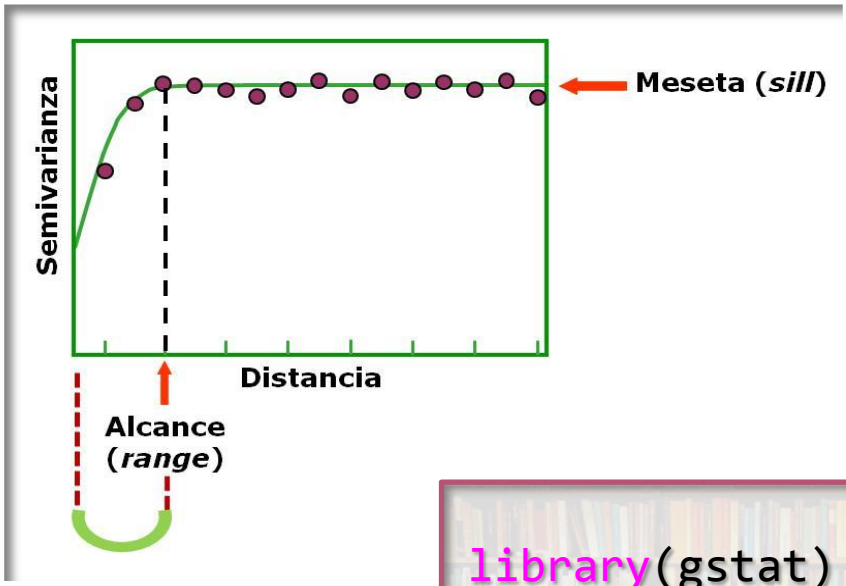
La semivarianza se define como:

$$I = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i + h) - z(x_i)]$$

donde

$z(x_i)$ y $z(x_i+h)$ son los **valores observados** de la variable regionalizada z en los puntos i e $i+h$, y

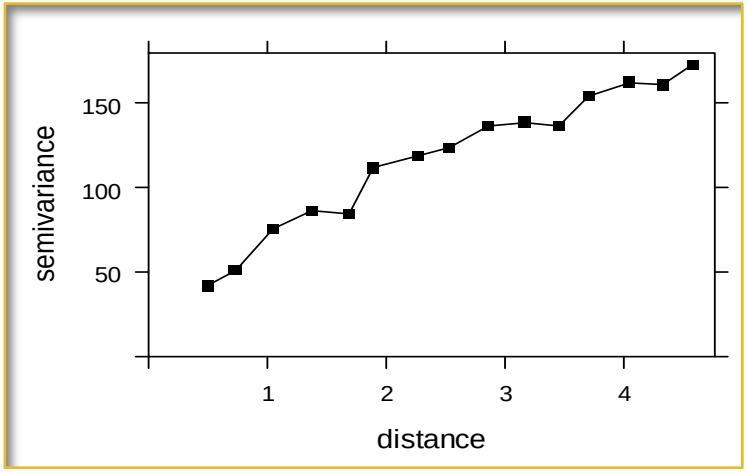
$N(h)$ es el **número de pares de puntos** separados entre sí por la distancia h




```
library(gstat)
```

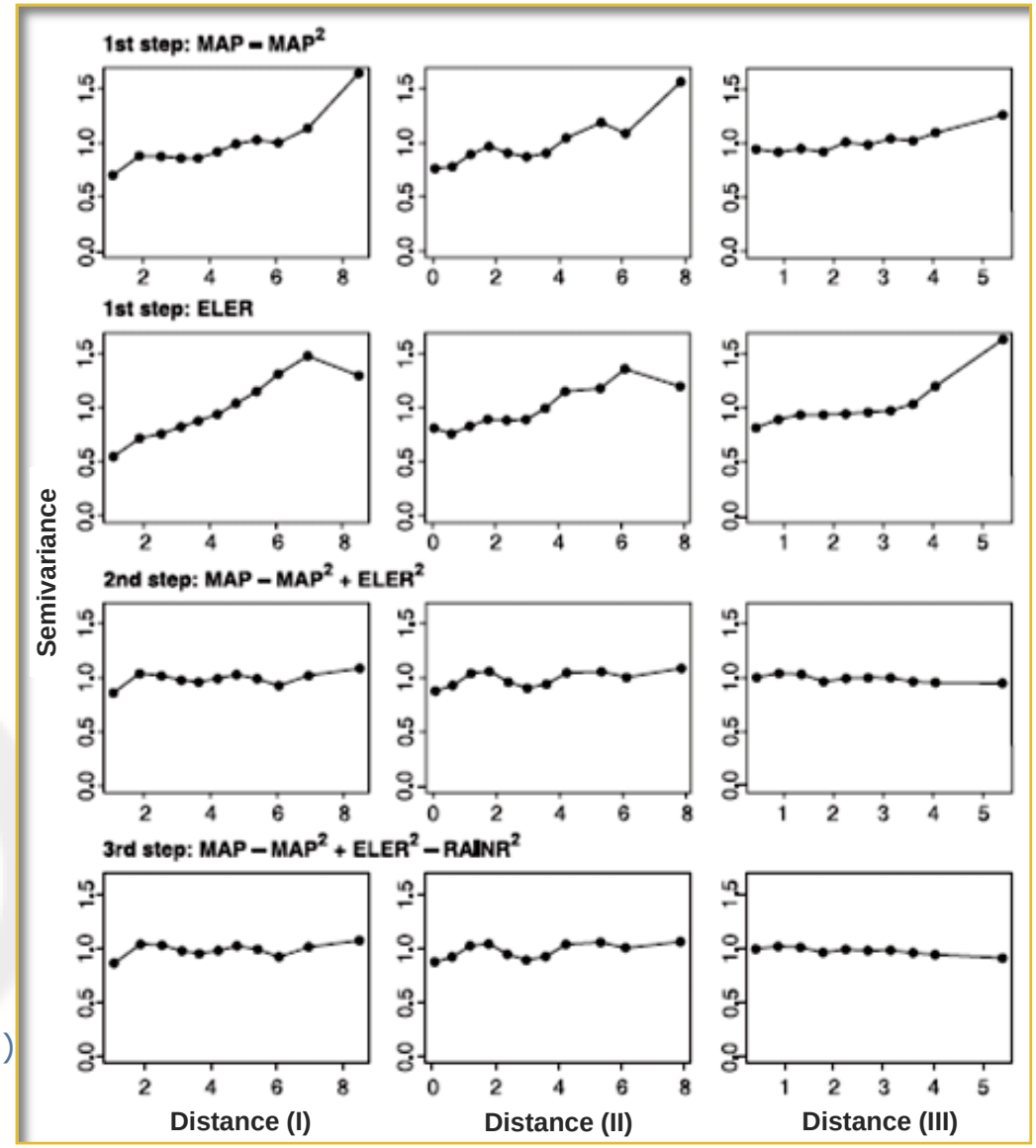
```
library(gstat)

plot(variogram(y ~ +1,
               locations = ~ lon+lat,
               data=mydata,
               width=0.3),
     col="black", lty=2, pch=15,
     type="b")
```



Semivariogramas

Variogramas de los residuales en modelos de regresión tras incluir las principales variables en cada uno de los pasos del proceso de selección.
[realizados con **S-Plus** 6.1 para Windows (Anónimo, 2002)]



Ferrer-Castán y Vetaas (2005)

Coeficientes de Moran I y correlogramas

Coeficiente de Moran *I*

$$I = \frac{\frac{1}{W}\sum_{hi}^{n} w_{hi}(y_h - y)(\bar{y}_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_h - \bar{y})^2}$$

donde

y_h e y_i: valores de la variable observada en los sitios h e i

w_{hi}: **pesos** que se dan a los pares de valores para una determinada **clase de distancia (1, 0)**

W = suma de los pesos = número de pares para una determinada clase de distancia

n = número total de puntos

Antes de computar los coeficientes:

- **Establecer clases de distancia** (espaciamiento irregular con igual número de pares de valores en cada clase de distancia)
- **Identificación de vecinos** en cada clase de distancia **(distancias euclídeas)**
- **Asignación de pesos** a las celdas vinculadas para crear un **matriz espacial de pesos** para cada una de las clases de distancia

Coeficiente de correlación lineal de Pearson (entre 2 variables)

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}$$

Coeficientes de Moran I y correlogramas

```
library(spdep)
-> library(sp)
-> library(Matrix)
-> library(lattice)
```

```
library(spdep)
coords <- as.matrix(cbind(lon, lat))
neigb1.nb <- dnearneigh(as.matrix(coords), d1=0.450, d2=1.470)
## Creación de la lista de vecinos (pares de puntos) para la clase 1...

summary(neigb1.nb)

is.symmetric.nb(neigb1.nb, verbose=NULL, force=FALSE) ## Para saber si el
objeto nb es simétrico o no. Si lo es, la relación entre los puntos i y j
es la misma que la relación entre los puntos j e i

d1.listw <- nb2listw(neigb1.nb) ## Asignación de pesos espaciales

moran.d1<- moran.test(residuals(fit1.lm), alternative="two.sided",
                      listw=d1.listw)
moran.d1 ## Displaya el coeficiente de Moran I y los tests asociados para
la primera clase de distancia
```

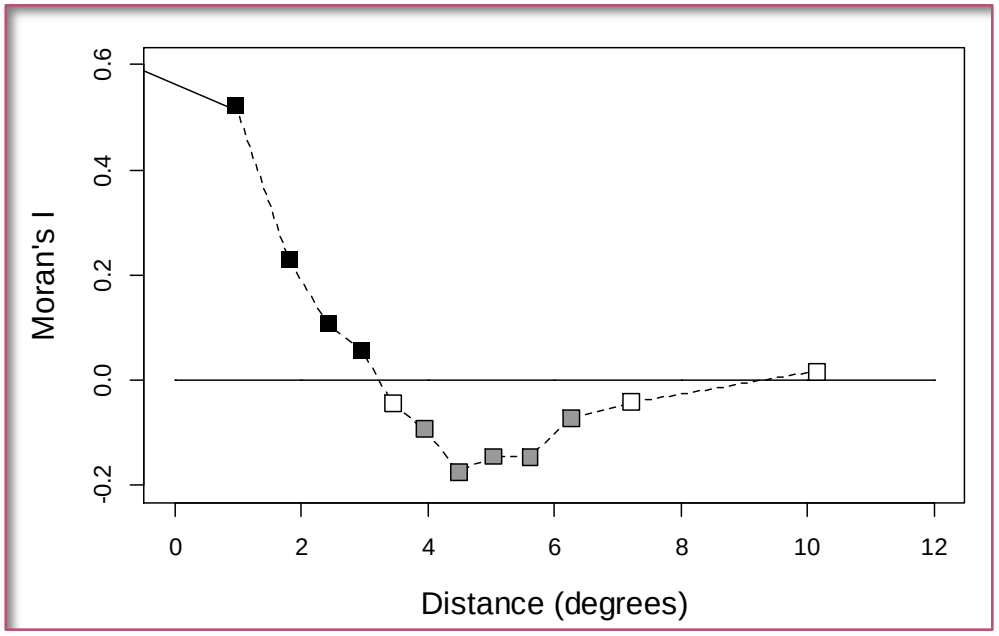
Coeficientes de Moran I y correlogramas

Clases de distancia

```
d1: 0.450-1.470
d2: 1.4701-2.159
d3: 2.1591-2.702
...
```

Puntos medios

```
mp1<-(0.450+1.470)/2 ## d1
mp2<-(1.4701+2.159)/2 ## d2
mp3<-(2.1591+2.702)/2 ## d3
```




```
x <- c(mp1, mp2, mp3...)
Moran.y <- c(0.524056228, 0.2293690675, -0.1068807343...) ## Se pueden
separar y diferenciar los valores positivos de los negativos, los
significativos de los que no lo son...

plot(x, Moran.y, ylim=c(-0.2, 0.6), xlim=c(0, 12), xlab="Distance (degrees)",
     ylab="Moran's I", cex.lab=1.3, type="l", lty=2)
axis(1, pos = 0, lty=1, tck=0, labels=F, xlim=c(0,10.5))
points(P.Moran.x, P.Moran.y, pch=15, cex=1.5)
...
```

Modelado espacialmente explícito

```
library(spdep)
-> library(sp)
-> library(Matrix)
-> library(lattice)
```

```
library(spdep)
```

Modelos espaciales autorregresivos

Fijan el proceso generador de errores y operan con las **matrices de pesos** que especifican la magnitud de las interacciones entre celdas vecinas

```
> spautolm(family = "SMA", method="eigen") ## Construye modelos autorregresivos simultáneos (SAR), condicionales (CAR) y basados en medias móviles (SMA). Los modelos SMA sólo están disponibles utilizando method="eigen“

> errorsarlm() ## Ajusta modelos de error SAR [idénticos a los que se obtienen con la función spautolm()]

> lagsarlm(type="lag") ## Ajusta modelos SAR de tipo "lag"

> lagsarlm(type="mixed") ## Para crear modelos mixtos SAR
```

Modelado espacialmente explícito

```
library(spdep)
-> library(sp)
-> library(Matrix)
-> library(lattice)
```

```
library(spdep)

coords <- as.matrix(cbind(lon, lat)) ## Para definir
coordinadas
neigb1.nb <- dnearneigh(as.matrix(coords), d1=0,45, d2=0,88) ## Creación
de la lista de vecinos (pares de puntos) para la clase 1

d1.listw <- nb2listw(neigb1.nb, style="W”) ## Asignación de pesos
espaciales
```

```
fit1.lm <- lm(y ~ x1+x2)
fit1.sma <- spautolm(fit1.lm, listw=d1.listw, family="SMA",
                     method="eigen", na.action="na.omit”) ## Con na.omit
se eliminan los NAs
```

Otras técnicas de evaluación

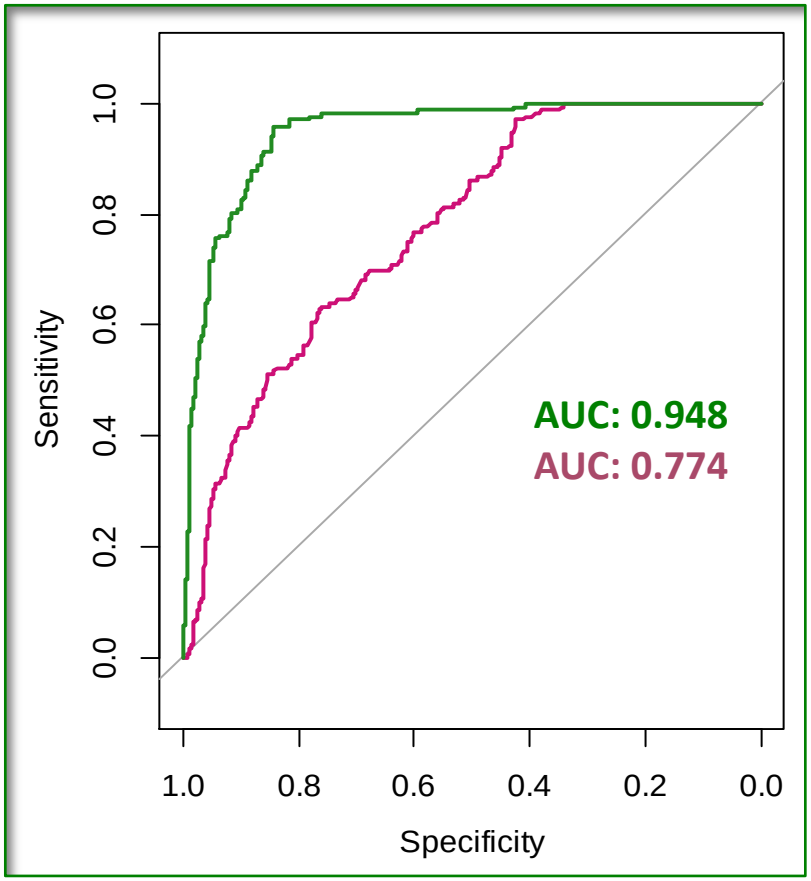
Curvas ROC

Curvas ROC

Matriz de confusión

	Presencia real	Ausencia real
Presencia predicha	**A** Verdadero positivo	**B** Falso negativo Error Comisión (sobrepredicción)
Ausencia predicha	**C** Falso negativo Error Omisión (subpredicción)	**D** Verdadero negativo

A/(A+C): SENSIBILIDAD (Fracción de verdaderos positivos)
C/(A+C): TASA OMISIÓN (Fracción de falsos positivos)
D/(D+B): ESPECIFICIDAD (Fracción de verdaderos negativos)



AUC (*Area under the curve*) (0.5-1)

[0.5, 0.6): Test malo
[0.6, 0.75): Test regular
[0.75, 0.9): Test bueno
[0.9, 0.97): Test muy bueno
[0.97, 1): Test excelente

Curvas ROC

```
library(pROC)
```

```
mydata <- read.table("C:/.../mydata.txt", header=TRUE, row.names=1)
attach(mydata)

fit1.glm <- glm(y ~ x, data=mydata, family="binomial")
coef(fit1.glm)

prob <- predict(fit1.glm,type=c("response"))
mydata$prob=prob
```

```
library(pROC)
g <- roc(y ~ prob, data=mydata)

plot(g, print.auc=TRUE)
```

Referencias

Anonymous (2002) *S-Plus 6.1 for Windows. Insightful Corporation,* Seattle, WA.

Ferrer-Castán, D. y Vetaas, O.R. (2005) Pteridophyte richness, climate and topography in the Iberian Peninsula: comparing spatial and nonspatial models of richness patterns. *Global Ecology and Biogeography*, **14,** 155-165.

Vetaas, O.R. y Ferrer-Castán, D. (2008) Patterns of woody plant richness in the Iberian Peninsula: environmental range and spatial scale. *Journal of Biogeography*, **35,** 1863-1878.

Paquetes de R utilizados

- **gstat**: *Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation*
- **lattice**: *Trellis Graphics for R*
- **Matrix**: *Sparse and Dense Matrix Classes and Methods*
- **mapdata**: *Extra Map Databases*
- **maps**: *Draw Geographical Maps*
- **pROC**: *Display and Analyze ROC Curves*
- **sp**: *Classes and Methods for Spatial Data*
- **spdep**: *Spatial Dependence: Weighting Schemes, Statistics and Models*

Disponibles en http://cran.es.r-project.org/

¡¡MUCHAS GRACIAS por su atención!!