

Informe sobre `rpart::kyphosis`
para Análisis de Datos 1 (MANADINE)

C. Carleos

N. Corral

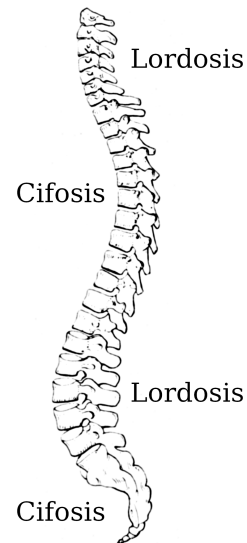
6 de abril de 2023

Capítulo 1

Introducción

1.1. Contexto

Se llama cifosis a las dos curvaturas que presenta la columna vertebral hacia afuera de forma natural, aunque a menudo se usa el término con sentido patológico, cuando la curvatura es demasiada. Los casos severos se tratan mediante cirugía correctiva. La foto¹ de abajo representa a un varón de 22 años antes de la operación.



Los datos analizados aquí pertenecen a un estudio que recoge información de 81 niños sometidos a tales intervenciones. Son un versión traducida (como se indica en el apéndice A.1) del *data frame* `kyphosis` contenido en el paquete `rpart`, el cual forma parte de la distribución básica de R (<http://r-project.org>). Más información se encuentra en el libro: John M. Chambers y Trevor J. Hastie (eds.) 1992. *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove, CA, EE.UU.

¹<https://es.wikipedia.org/wiki/Cifosis>

1.2. Variables

Se recogen los datos en un *data frame* llamado `cifosis`, que contiene las siguientes columnas:

Cifosis – Toma valores `presente` o `ausente` en función de si el individuo de la muestra presenta o no cifosis tras haber sido sometido a cirugía correctiva.

Edad – Edad en meses.

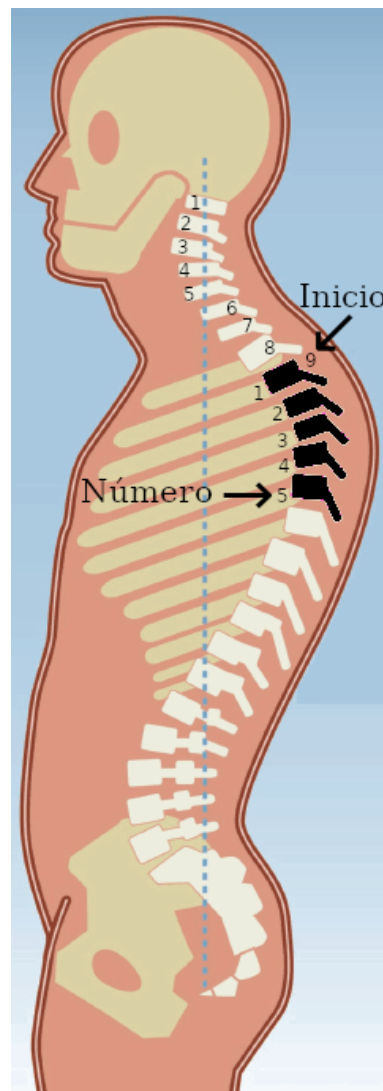
Número – Cuántas vértebras hay involucradas.²

Inicio – Posición (desde lo alto) de la primera vértebra operada.

1.3. Objetivos

En el trabajo se plantean los siguientes objetivos:

- Analizar la relación entre las características de los pacientes (Edad, Número e Inicio) y el resultado del tratamiento quirúrgico.
- Obtener un modelo para predecir la eficacia de la cirugía (Cifosis) a partir de dichas características.
- Describir la calidad del modelo.



²Ilustración adaptada de <https://centrocorporesano.es/wp-content/uploads/2019/09/cifosis-al-detalle.png>

Capítulo 2

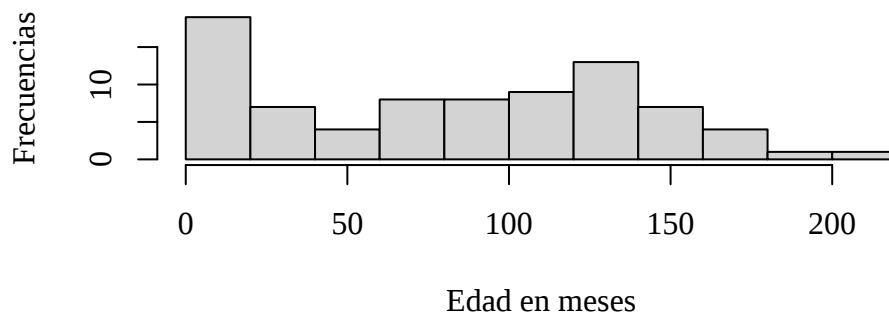
Descriptiva

2.1. Univariante

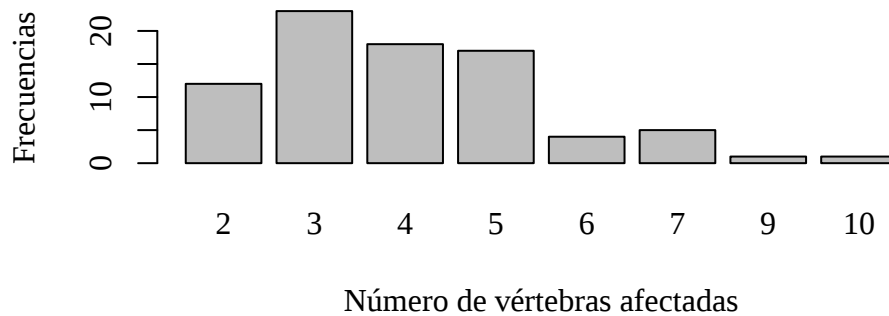
Cifosis – Muchos más casos de cirugía exitosa (**ausente**). Frecuencias:

	ausente	presente
absolutas	64	17
relat. (%)	79	21

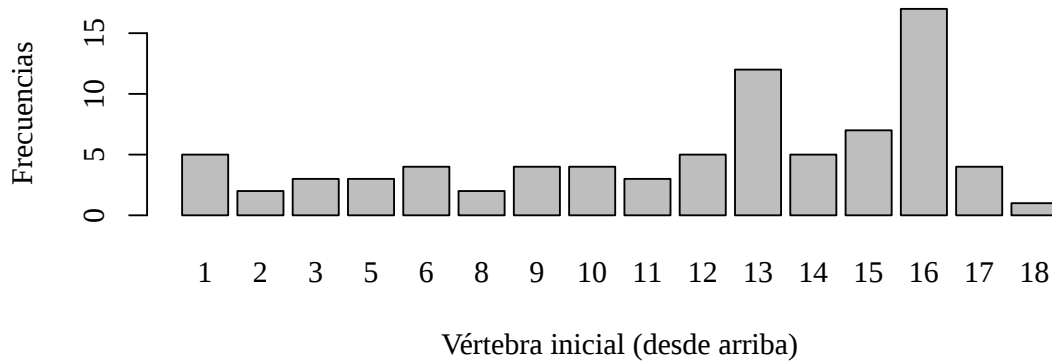
Edad – Menores de 18 años, con acumulación de bebés.



Número – En general, de 2 a 5 vértebras afectadas.



Inicio – Casos desde la 1 (la más alta), pero mayor concentración entre la 13 y la 16.

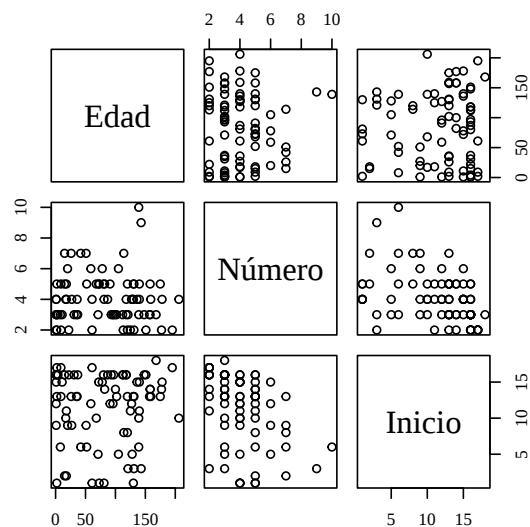


2.2. Bivariante

2.2.1. Entre explicativas

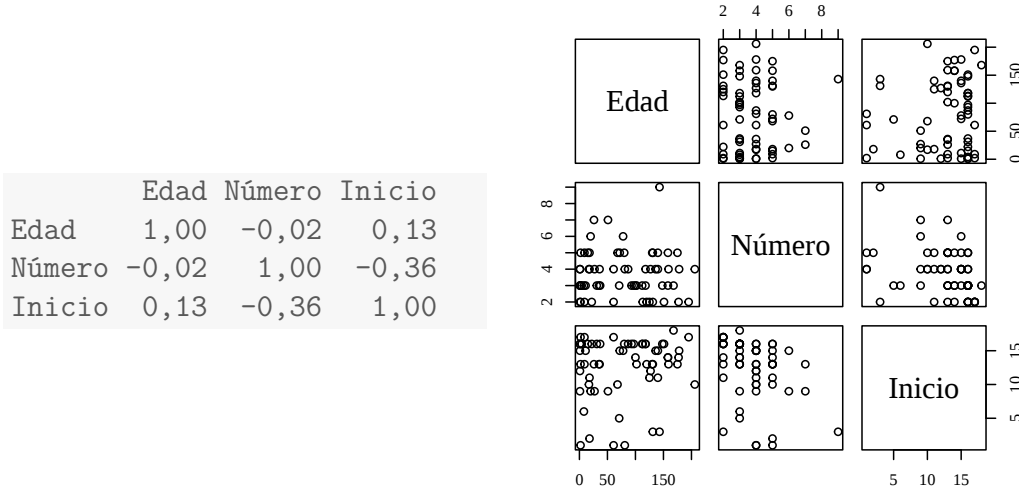
Correlación negativa entre **Número** e **Inicio**, como cabe esperar (si se empieza más abajo, hay menos vértebras disponibles). Correlación nula en el resto de parejas de explicativas, sin que haya relación no lineal. Matriz de correlaciones:

	Edad	Número	Inicio
Edad	1,00	-0,02	0,06
Número	-0,02	1,00	-0,43
Inicio	0,06	-0,43	1,00

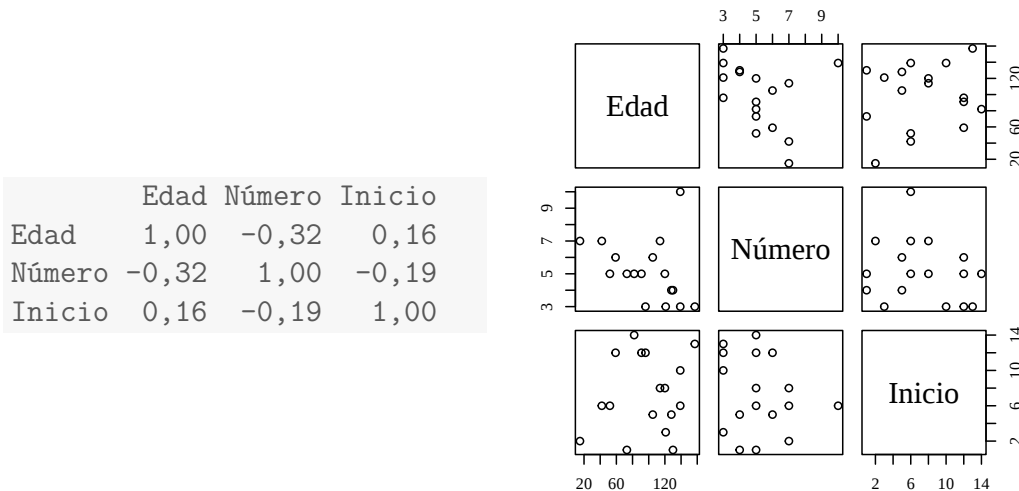


Al analizar las correlaciones intra-grupo se aprecia que cuando Cifosis == presente hay cierta correlación negativa entre Número y Edad, por lo que convendría probar un análisis discriminante cuadrático, además del lineal.

Cifosis==ausente



Cifosis==presente

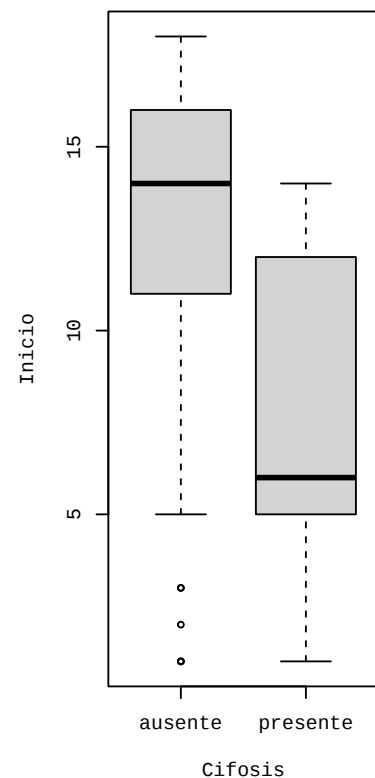
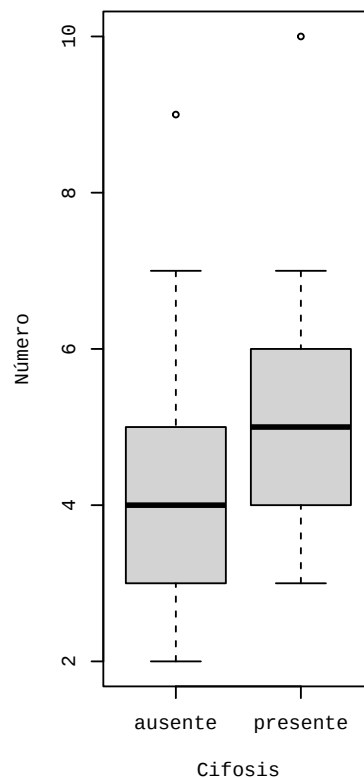
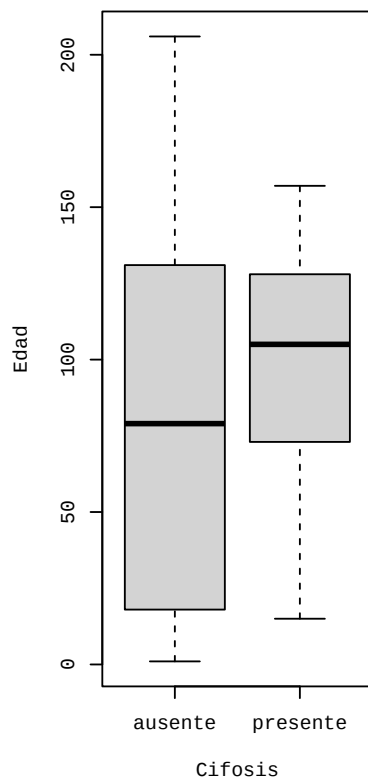


2.2.2. Entre explicativas y respuesta

La variable respuesta está asociada a cada variable explicativa, sobre todo con Inicio. En gran medida, $\text{Inicio} > 12 \implies \text{ausente}$, $\text{Inicio} < 11 \implies \text{presente}$. De hecho, con este último umbral se consigue más de un 70 % de acierto:

```
prop.table (table (real=cifosis$Cifosis, pred=cifosis$Inicio<11), "real")
```

```
##           pred
## real      FALSE      TRUE
##  ausente  0,7656250 0,2343750
##  presente 0,2941176 0,7058824
```



Capítulo 3

Modelo predictivo

Usaremos la técnica de análisis discriminante por ser la única supervisada vista en esta asignatura. En R se calcula así:

```
library (MASS)
(modelo <- lda (Cifosis ~ ., cifosis))

## Call:
## lda(Cifosis ~ ., data = cifosis)
##
## Prior probabilities of groups:
##   ausente   presente
## 0,7901235 0,2098765
##
## Group means:
##           Edad  Número  Inicio
## ausente  79,89062 3,750000 12,609375
## presente 97,82353 5,176471  7,294118
##
## Coefficients of linear discriminants:
##           LD1
## Edad      0,005910971
## Número    0,291501797
## Inicio   -0,170496626
```

El modelo construye el discriminante lineal $0,006 \cdot \text{Edad} + 0,292 \cdot \text{Número} + -0,17 \cdot \text{Inicio}$. Los signos de los coeficientes indican que la “falta de eficacia” de la cirugía tiende a incrementarse:

- con la Edad,
- con el Número de vértebras con problemas,

- cuanto más alta esté la primera vértebra afectada.

Para cuantificar la importancia de las variables se puede aplicar el discriminante lineal a los datos tipificados (fig. A.2, pág. 12), lo que produce el modelo $0,343 \cdot \text{Edad} + 0,472 \cdot \text{Número} + -0,833 \cdot \text{Inicio}$, que refleja lo visto antes: **Inicio** es la variable más asociada a la respuesta.

El modelo¹ clasifica como **presente** para valores mayores que 1,2 aproximadamente. Dicho punto de corte 1,2 se ha buscado a ojo para equilibrar las probabilidades de acierto (o error) en ambos grupos:

```
pmodelo <- predict (modelo)
table (pmodelo$class, pmodelo$x > 1.2)

##
##          FALSE TRUE
## ausente      67    0
## presente      0   14
```

La clasificación mostrada no tiene errores, lo que sugiere un sobreajuste. Para obtener una idea correcta de la calidad del modelo, se aplica validación cruzada dejando uno fuera (*jackknife*), que resulta en la siguiente matriz de confusión:

absolutas:

real	pred	
	ausente	presente
ausente	58	6
presente	10	7

porcentajes:

real	pred	
	ausente	presente
ausente	91	9
presente	59	41

Que para los **presente** haya más errores que aciertos puede llevar a poner en duda la conveniencia de usar probabilidades a priori estimadas a partir de la muestra. El análisis hecho con aprioris equiprobables da

absolutas:

real	pred	
	ausente	presente
ausente	51	13
presente	6	11

porcentajes:

real	pred	
	ausente	presente
ausente	80	20
presente	35	65

y podría ser preferido un resultado así, más equilibrado entre los dos grupos. Como desconocemos posibles cuantificaciones de pérdidas por error, mantendremos el modelo calculado en primer lugar.

¹Con las variables originales, no las tipificadas.

Un discriminante cuadrático mejora algo los resultados (fig. A.2, pág. 12) cuando los aprioris se estiman de la muestra, pero no parece una mejora relevante para justificar un modelo más complejo y mucho menos interpretable. Con aprioris equiprobables, no hay mejora ninguna, por lo que decidimos quedarnos con el discriminante lineal.

3.1. Conclusiones

- La variable más relevante para predecir **Cifosis** es **Inicio**. Mediante ella sola se consigue más del 70 % de aciertos.
- Se ha llegado a la fórmula

$$F = 0,006 \cdot \text{Edad} + 0,292 \cdot \text{Número} + -0,17 \cdot \text{Inicio}$$

que clasifica como **presente** si $F > 1,2$ aproximadamente.

- Dicha fórmula consigue un 91 % de aciertos en **ausente** pero sólo un 41 % entre los **presente**; en total, un 80 %.
- Se podrían especificar probabilidades a priori para equilibrar los porcentajes de acierto en ambos grupos.

Apéndice A

Anexos

A.1. Traducción

```
cifosis <- rpart::kyphosis
names(cifosis) <- c("Cifosis", "Edad", "Número", "Inicio")
levels(cifosis$Cifosis) <- c("ausente", "presente")
head(cifosis)
```

Más traducciones:

```
options (OutDec = ",")
histograma <- function (x, ...)
  hist(x, ..., main="", ylab = "Frecuencias")
barras <- function (x, ...) barplot (table (x), ..., main="",
                                     ylab = "Frecuencias")
```

A.2. Discriminante

Figura A.1: Discriminante lineal con variables originales.

```
library (MASS)
(modelo <- lda (Cifosis ~ ., cifosis))

## Call:
## lda(Cifosis ~ ., data = cifosis)
##
## Prior probabilities of groups:
##   ausente   presente
## 0,7901235 0,2098765
##
## Group means:
##           Edad   Número   Inicio
## ausente  79,89062 3,750000 12,609375
## presente 97,82353 5,176471  7,294118
##
## Coefficients of linear discriminants:
##           LD1
## Edad      0,005910971
## Número    0,291501797
## Inicio   -0,170496626
```

Figura A.2: Discriminante lineal con variables tipificadas.

```
library (MASS)
(modelo <- lda (Cifosis ~ ., cifosis))

## Call:
## lda(Cifosis ~ ., data = cifosis)
##
## Prior probabilities of groups:
##   ausente   presente
## 0,7901235 0,2098765
##
## Group means:
##           Edad  Número  Inicio
## ausente  79,89062 3,750000 12,609375
## presente 97,82353 5,176471  7,294118
##
## Coefficients of linear discriminants:
##           LD1
## Edad      0,005910971
## Número    0,291501797
## Inicio   -0,170496626
```

Figura A.3: Discriminante cuadrático.

Aprioris MUESTRALES
absolutas:

	pred	
real	ausente	presente
ausente	59	5
presente	10	7

Aprioris EQUIPROB.
absolutas:

	pred	
real	ausente	presente
ausente	51	13
presente	6	11

porcentajes:

	pred	
real	ausente	presente
ausente	92	8
presente	59	41

porcentajes:

	pred	
real	ausente	presente
ausente	80	20
presente	35	65

A.3. Código R de los análisis

Descriptiva univariante:

```
rbind (absolutas = (t <- table(cifosis$Cifosis)),
      "relat. (%)" = round(100*prop.table(t)))
par (mar = c(4,4,1,1))
histograma(cifosis$Edad, xlab="Edad en meses")
par (mar = c(4,4,1,1))
barras(cifosis$Número, xlab="Número de vértebras afectadas")
barras(cifosis$Inicio, xlab="Vértebra inicial (desde arriba)")
```

Descriptiva bivalente:

```
round (cor (subset (cifosis, , -Cifosis)), 2)
par (mar = c(0,0,0,0))
pairs (subset (cifosis, , -Cifosis))
par(mfrow=c(1,3))
boxplot (Edad ~ Cifosis, cifosis)
boxplot (Número ~ Cifosis, cifosis)
boxplot (Inicio ~ Cifosis, cifosis)
par(mfrow=c(1,1))
```

Discriminante lineal:

```
library (MASS)
(modelo <- lda (Cifosis ~ ., cifosis))
tabla <- table (real=cifosis$Cifosis, pred=lda(Cifosis~.,cifosis,CV=TRUE)$class)
cat ("absolutas:")
tabla
cat ("porcentajes:")
round (100 * prop.table (tabla, "real"))
```

Discriminante cuadrático:

```
cat ("Aprioris MUESTRALES")
tabla <- table (real=cifosis$Cifosis,
               pred=qda(Cifosis~.,cifosis,CV=TRUE)$class)
cat ("absolutas:")
tabla
cat ("porcentajes:")
round (100 * prop.table (tabla, "real"))
cat ("Aprioris EQUIPROB.")
```

```
tabla <- table (real=cifosis$Cifosis,  
               pred=qda(Cifosis~.,cifosis,CV=TRUE,prior=c(.5,.5))$class)  
cat ("absolutas:")  
tabla  
cat ("porcentajes:")  
round (100 * prop.table (tabla, "real"))
```