MR - Trabajo

Alicia Losada | alicia.losada.sanchez@udc.es María Cardoso | m.cardoso@udc.es
Nicolás Muñiz | nicolas.muniz@udc.es

11/12/2024

Regresión Lineal Múltiple

1. Análisis descriptivo

2. Modelo matemático

$$\mathbb{E}(\vec{Y}|\boldsymbol{X}) = \beta_0 + \sum_{i=1}^{n} \beta_i X_{ij} \tag{1}$$

Regresión Logística

- Antes de empezar, cargamos los datos *Oro.rda*

```
load("Datos/Oro.rda")
attach(Oro)
explicativas.oro <- Oro[,1:3]     # Almacenamos las explicativas
respuesta.oro <- Proximidad       # Almacenamos la variable de respuesta
```

1. Análisis descriptivo

Para el análisis descriptivo de las variables podemos comenzar con una visión general de las variables mediante las funciones `str()` y `summary()`.

```
str(Oro)
```

```
## 'data.frame':    64 obs. of  4 variables:
##  $ As        : num  6.77 15.03 6.43 0.1 0.1 ...
##  $ Sb        : num  3.08 6.15 2.35 0.3 0.3 9.62 0.51 3.71 4.32 0.8 ...
##  $ Corredor  : int  1 1 1 0 0 1 0 1 0 0 ...
##  $ Proximidad: int  1 1 1 0 0 1 0 1 0 0 ...
```

La salida de `str()` nos dice que los datos constan de 64 observaciones de 4 variables:

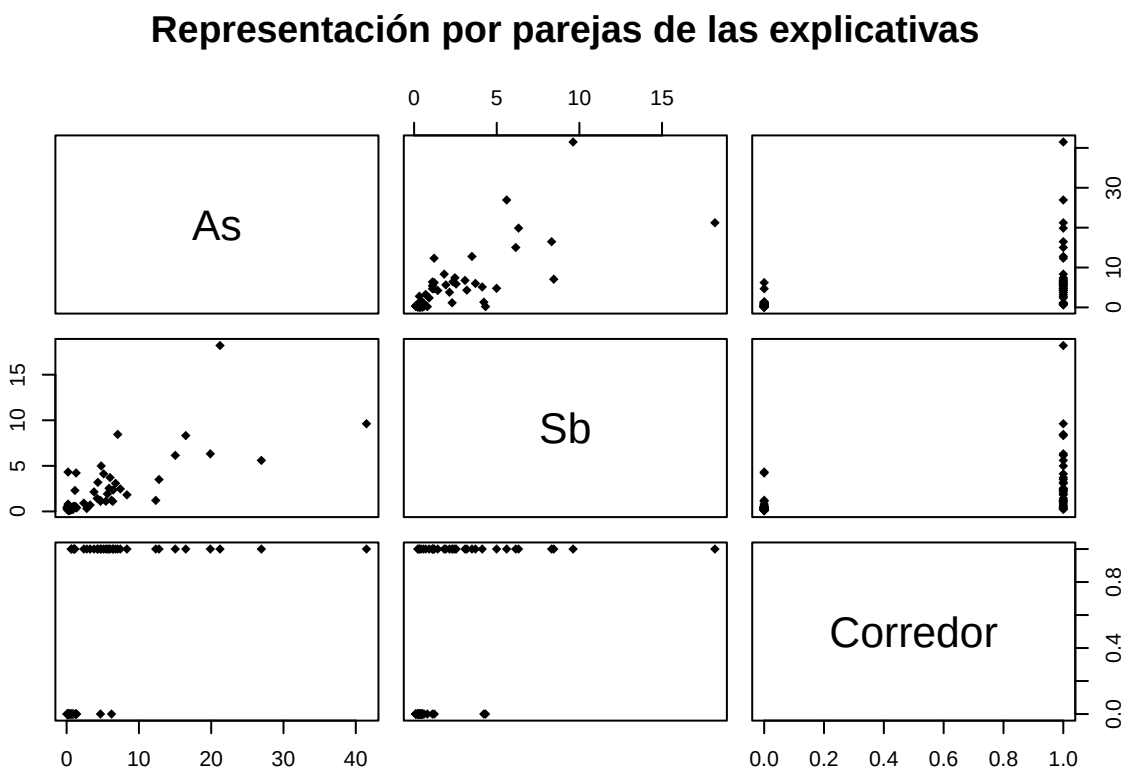
- `As`: Nivel de concentración de arsénico en la muestra de agua. (numérica)
- `Sb`: Nivel de concentración de antimonio en la muestra de agua. (numérica)
- `Corredor`: Variable binaria indicando si la zona muestreada está (1) o no está (0) en alguno de los corredores delimitados por las lineas sobre el mapa. (categórica)

Más la variable de respuesta `Proximidad`, que toma los valores 1 o 0 según que el depósito esté próximo o esté muy lejano al lugar.

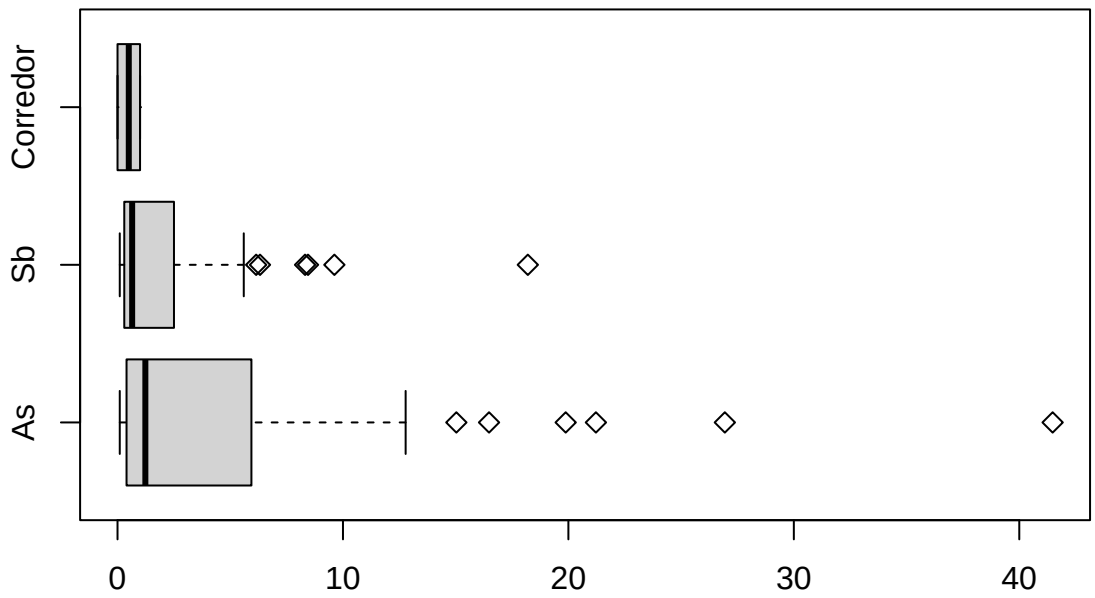
```
summary(Oro)
```

```
##        As               Sb            Corredor     Proximidad
##  Min.   : 0.100   Min.   : 0.100   Min.   :0.0   Min.   :0.0000
##  1st Qu.: 0.400   1st Qu.: 0.300   1st Qu.:0.0   1st Qu.:0.0000
##  Median : 1.235   Median : 0.650   Median :0.5   Median :0.0000
##  Mean   : 4.645   Mean   : 2.039   Mean   :0.5   Mean   :0.4375
##  3rd Qu.: 5.905   3rd Qu.: 2.487   3rd Qu.:1.0   3rd Qu.:1.0000
##  Max.   :41.480   Max.   :18.200   Max.   :1.0   Max.   :1.0000
```

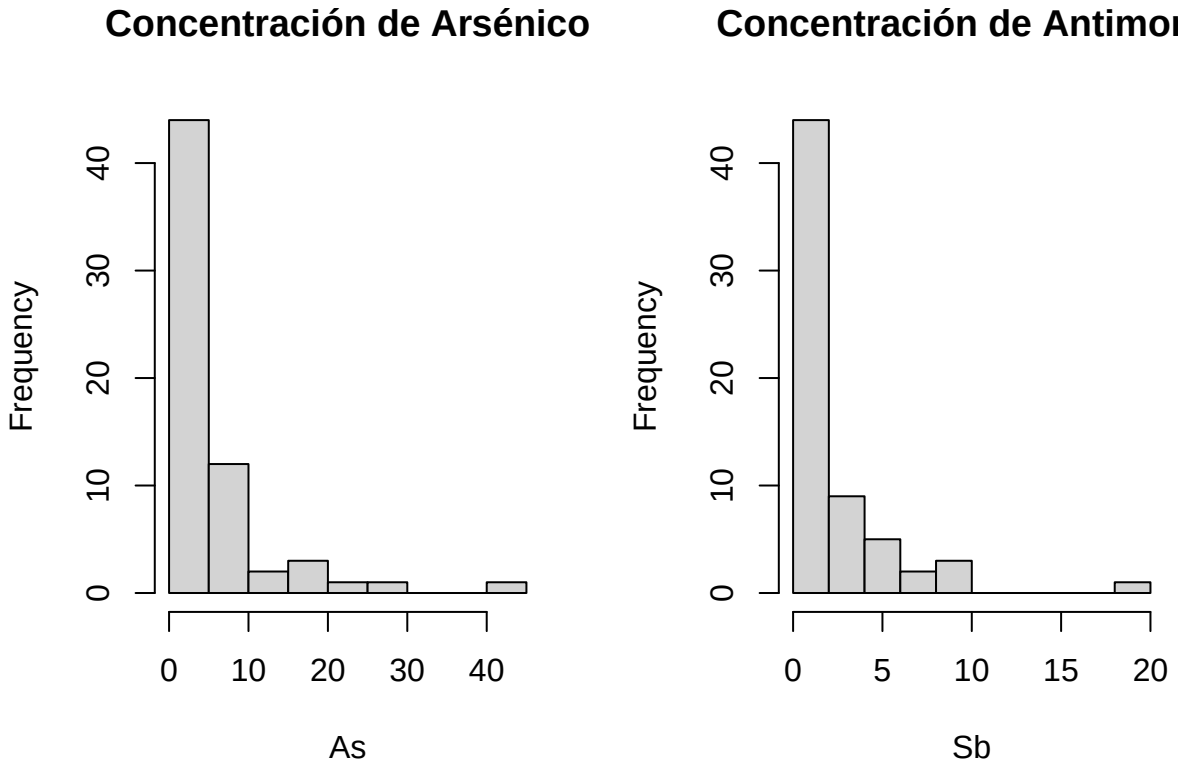
```
plot(explicativas.oro, pch=18,
     main="Representación por parejas de las explicativas")
```



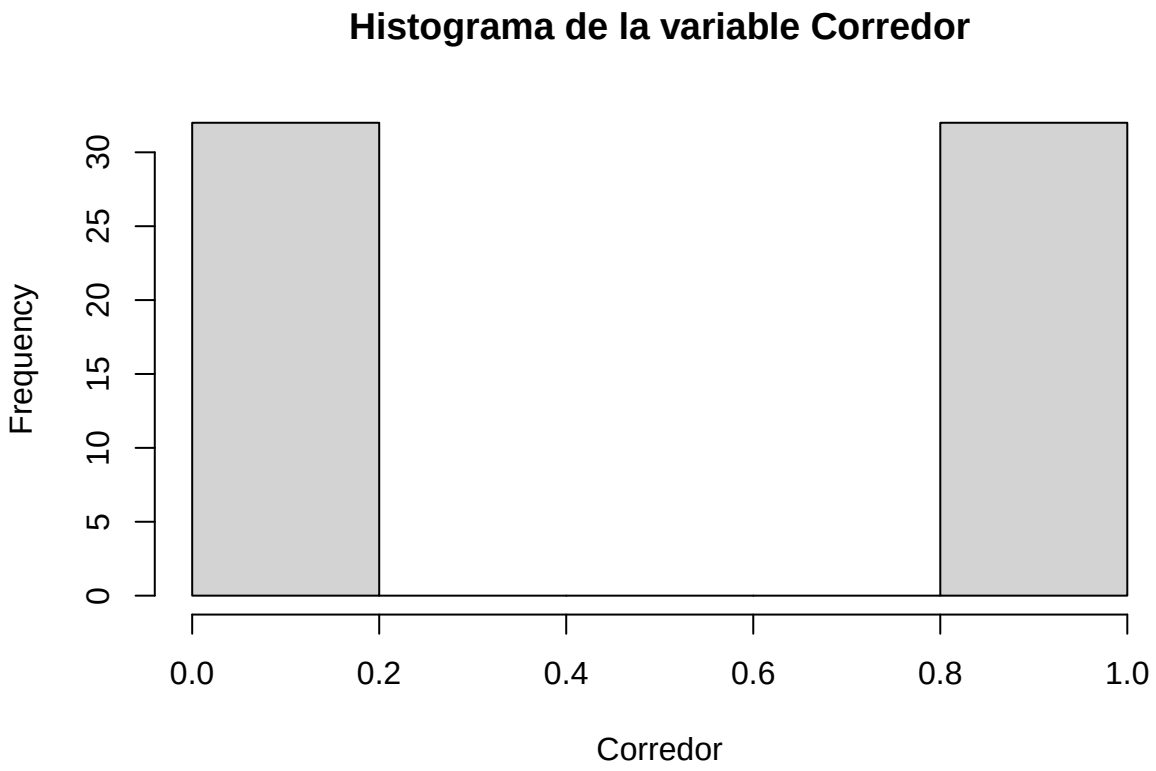

```
boxplot(explicativas.oro, horizontal=T, pch=5,
        main="Diagrama de cajas de las explicativas")
```




```
old.par <- par(mfrow=c(1,2))
hist(As, main="Concentración de Arsénico")
hist(Sb, main="Concentración de Antimonio")
```



```r
par(old.par)
hist(Corredor, main="Histograma de la variable Corredor")
```



2. Modelo matemático

Dado que la variable de respuesta, `Proximidad`, es binaria (0 o 1), deberemos de elegir un modelo que tenga esto en cuenta. En nuestro caso hemos elegido una transformación del modelo lineal, definida por la distribución logística de la ecuación 2

$$F(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \tag{2}$$

Por tanto, nuestro modelo logístico quedaría de la forma

$$\mathbb{E}(Y|\vec{X}_i) = p_i = \mathbb{P}(Y=1|\vec{X}_i) = \frac{1}{1+e^{-\eta}} \tag{3}$$

tal que $\eta = \vec{\beta}^t \vec{X}_i$. Además,

$$1 - p_i = \mathbb{P}(Y=0|\vec{X}_i) = 1 - \frac{1}{1+e^{-\eta}} = \frac{e^{-\eta}}{1+e^{-\eta}} \tag{4}$$

3. Interpretación del modelo

Para una mejor interpretación del modelo, podemos definir el $\mathbf{odds}_i$ de manera que

$$odds_i = odds(Y|\vec{X}_i) = \frac{p_i}{1-p_i} = e^{\eta} = e^{\vec{\beta}^t \vec{X}_i} = e^{\beta_0} e^{\beta_1 X_{i1}} \cdots e^{\beta_k X_{ik}} = e^{\beta_0} \prod_{j=1}^{k} e^{\beta_j X_{ij}} \ , \ 1 \leq i \leq n \qquad (5)$$

Este es un modelo multiplicativo, en el cual e^{β_0} es la respuesta cuando $\vec{X}_i = \vec{0}$, mientras que e^{β_j}, para $1 \leq j \leq k$, es el incremento multiplicativo $(e^{\beta_j})^l$ en el odds para algún incremento l en X_j

También podemos expresar el modelo aplicando logaritmos a la ecuación 5, de manera que

$$\ln(\frac{p_i}{1-p_i}) = \eta = \vec{\beta}^t \vec{X}_i \qquad (6)$$

Los cuales denominaremos como $\mathbf{logit}_i$. Estos logits son interpretables mucho más fácilmente, aunque debido a que

4. Inferencia

5. Bondad del ajuste