# Modelos Lineales Generalizados: aplicaciones en R

# Modelos Lineales Generalizados: un enfoque aplicado

Ana M. Bianco Jemina García
(anambianco@gmail.com) (jeminagarcia@gmail.com )

## 3. El problema de Clasificación y Regresión Logística

# El problema de clasificación

Consideremos una variable categórica $Y$ que toma valores 0 y 1, que puede indicar la pertenencia a una categoría o a una clase o a un estado, por ejemplo sano o enfermo y se quiere predecir el estado en función de otras variables $(X_1, ..., X_p)$ por ejemplo: peso, edad, nivel de colesterol, nivel de glucosa en sangre, presión sanguínea.

- Spam o no Spam: queremos clasificar un correo como spam o no de acuerdo a un conjunto de características del correo: presencia de ciertas palabras, país de origen, etc.
- Pago de trajeta de crédito: un banco quiere predecir si un cliente incurrirá en un impago de su tarjeta en base a algunas variables como edad, salario, impagos en los últimos 3 meses, etc.
- Clasificación de hongos: queremos clasificar dos especies de hongos de acuerod a sus características morfológicas.

# Clasificador

- Información disponible $X = (X_1, X_2, \ldots, X_p) \in \mathcal{X}$.
- Posibles *etiquetas*. Caso binario $\mathcal{Y} = \{0, 1\}$
- Posibles *etiquetas*. Caso general $\mathcal{Y} = \{y_1, \ldots, y_k\}$
- Clasificador: Regla que asigna a $x \in \mathcal{X}$ un posible valor $y \in \mathcal{Y}$.

# Diccionario - Wasserman & Hastie

$$\mathbf{X} = (X_1, X_2, \ldots, X_p) \in \mathcal{X} \quad Y \in \{0, 1\}\,, \quad Y \in \mathcal{Y}$$

| Estadística | Computer Science | Significado |
|---|---|---|
| Clasificación | Aprendizaje Supervisado | Predecir $Y$ categórica con $\mathbf{X}$ |
| Covariables | Features | las $X_i'$s |
| Clasificador | Hypothesis | map $h = \mathcal{X} \to \mathcal{Y}$ |
| data | training sample | $(\mathbf{X_1}, Y_1), (\mathbf{X_2}, Y_2), \ldots (\mathbf{X_n}, Y_n)$ |

# Clasificación: Marco Teórico

- $X \in \mathcal{X}$, $Y \in \mathcal{Y}$.
- $(X, Y)$ vector aleatorio, con conjunta $F_{XY}$.
- Clasificador: Regla (de clasificación) que asigna a cada $x \in \mathcal{X}$ un elemento $y \in \mathcal{Y}$

$$\text{Clasificador} \quad H : \mathcal{X} \to \mathcal{Y}$$

- Error de Clasificación Medio (verdadero - poblacional) del clasificador $H$

$$L(H) = \mathbb{P}(H(X) \neq Y)$$

- Objetivo (teórico): Encontrar $H$ que minimice el error medio de clasificación.

# $H^{opt}$ Optimo: Regla de Bayes - Caso binario

$$H^{opt}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y=1 \mid X=x) > \mathbb{P}(Y=0 \mid X=x), \\ 0 & \text{si} \quad \mathbb{P}(Y=0 \mid X=x) > \mathbb{P}(Y=1 \mid X=x). \end{cases}$$

# $H^{opt}$ Optimo: Regla de Bayes - Caso binario

$$H^{opt}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y=1 \mid X=x) > \mathbb{P}(Y=0 \mid X=x), \\ 0 & \text{si} \quad \mathbb{P}(Y=0 \mid X=x) > \mathbb{P}(Y=1 \mid X=x). \end{cases}$$

**Teorema:** Para todo $H$

$$L(H^{opt}) = P\left(H^{opt}(X) \neq Y\right) \leq P\left(H(X) \neq Y\right) = L(H)\,.$$

# Error de Clasificación Empírico de $H$

- $X \in \mathcal{X}$, $Y \in \mathcal{Y}$.
- $(X, Y)$ vector aleatorio, con conjunta $F_{XY}$.
- Clasificador: Regla (de clasificación) que asigna a cada $x \in \mathcal{X}$ un elemento $y \in \mathcal{Y}$

$$\text{Clasificador} \quad H : \mathcal{X} \to \mathcal{Y}$$

- Error de Clasificación Medio (verdadero - poblacional) del clasificador $H$

$$L(H) = \mathbb{P}(H(X) \neq Y)$$

# Error de Clasificación Empírico de $H$

- $X \in \mathcal{X}$, $Y \in \mathcal{Y}$.
- $(X, Y)$ vector aleatorio, con conjunta $F_{XY}$.
- Clasificador: Regla (de clasificación) que asigna a cada $x \in \mathcal{X}$ un elemento $y \in \mathcal{Y}$

$$\text{Clasificador} \quad H : \mathcal{X} \to \mathcal{Y}$$

- Error de Clasificación Medio (verdadero - poblacional) del clasificador $H$

$$L(H) = \mathbb{P}(H(X) \neq Y)$$

- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Error de Clasificación Empírico del clasificador $H$: proporción de pares mal clasificados según $H$.

$$\widehat{L}_n(H) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{H(x_i) \neq y_i}$$

# Error de Clasificación Empírico de $\hat{H}_n$

- $X \in \mathcal{X}$, $Y \in \mathcal{Y}$.
- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- $\hat{H}_n$: Procedimiento contruido con los datos.

$$\hat{H}_n : \mathcal{X} \to \mathcal{Y}$$

- Error de Clasificación Empírico del clasificador $\hat{H}_n$:

$$\widehat{L}_n(\hat{H}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\hat{H}_n(x_i) \neq y_i}$$

# Error de Clasificación Empírico de $\hat{H}_n$

- $X \in \mathcal{X}$, $Y \in \mathcal{Y}$.
- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- $\hat{H}_n$: Procedimiento contruido con los datos.

$$\hat{H}_n : \mathcal{X} \to \mathcal{Y}$$

- Error de Clasificación Empírico del clasificador $\hat{H}_n$:

$$\widehat{L}_n(\hat{H}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\hat{H}_n(x_i) \neq y_i}$$

Cuidado: Overfitting!

- Error de Clasificación leave one out (cross – validation) de la regla:

$$\text{CV} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\hat{H}_n^{(-i)}(x_i) \neq y_i}.$$

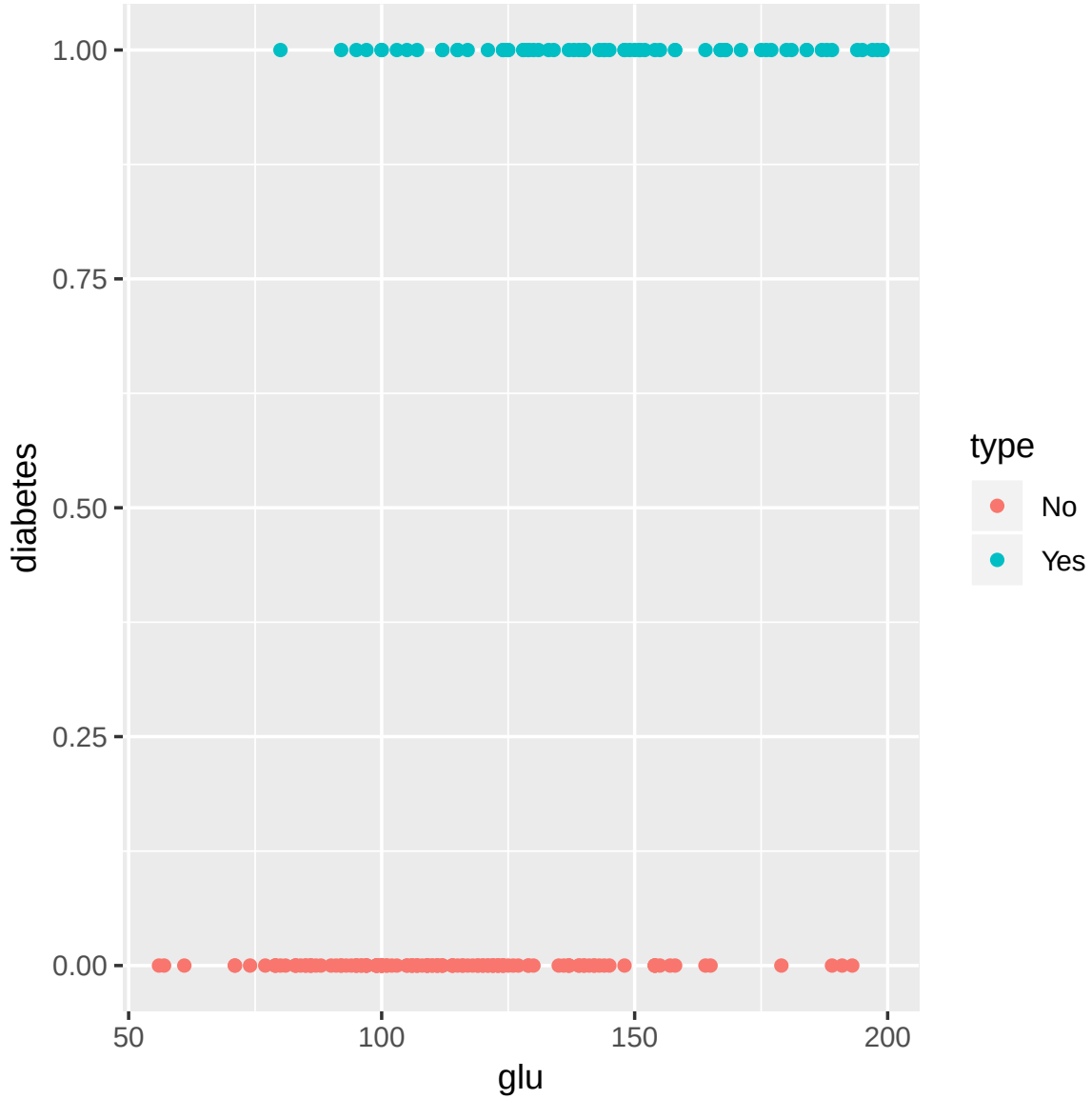
# Vayamos un poco para atrás....

- $x$ v.a. $\in \mathcal{X}$
- Posibles *etiquetas*. Caso binario $\mathcal{Y} = \{0, 1\}$
- Clasificador: Regla que asigna a cada $x \in \mathcal{X}$ un elemento $y \in \mathcal{Y}$
- $H^{opt}$ Optimo: Regla de Bayes - Caso binario

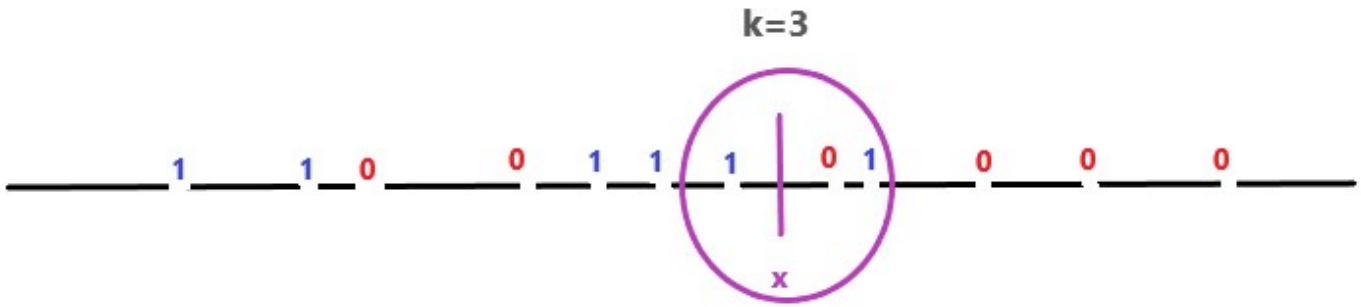
$$H^{opt}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y = 1 \mid X = x) > \mathbb{P}(Y = 0 \mid X = x), \\ 0 & \text{si} \quad \mathbb{P}(Y = 0 \mid X = x) > \mathbb{P}(Y = 1 \mid X = x). \end{cases}$$

¿Cómo podríamos estimar $\mathbb{P}(Y = 1 \mid X = x)$ y $\mathbb{P}(Y = 0 \mid X = x)$?

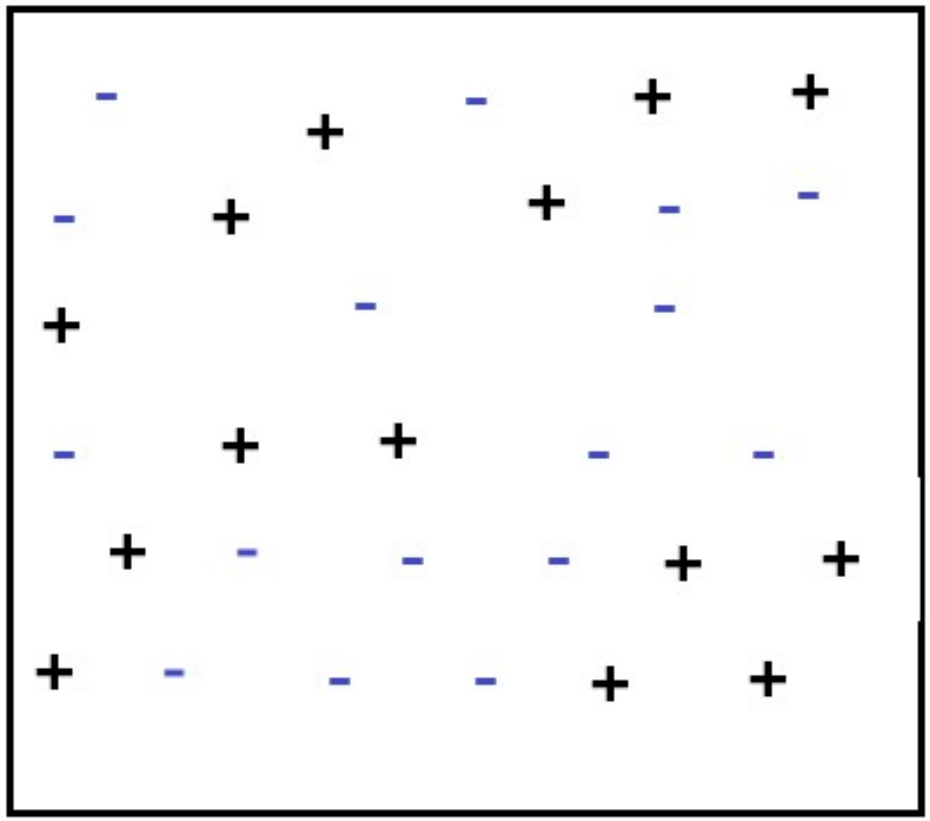
# Volvamos a los datos PIMA



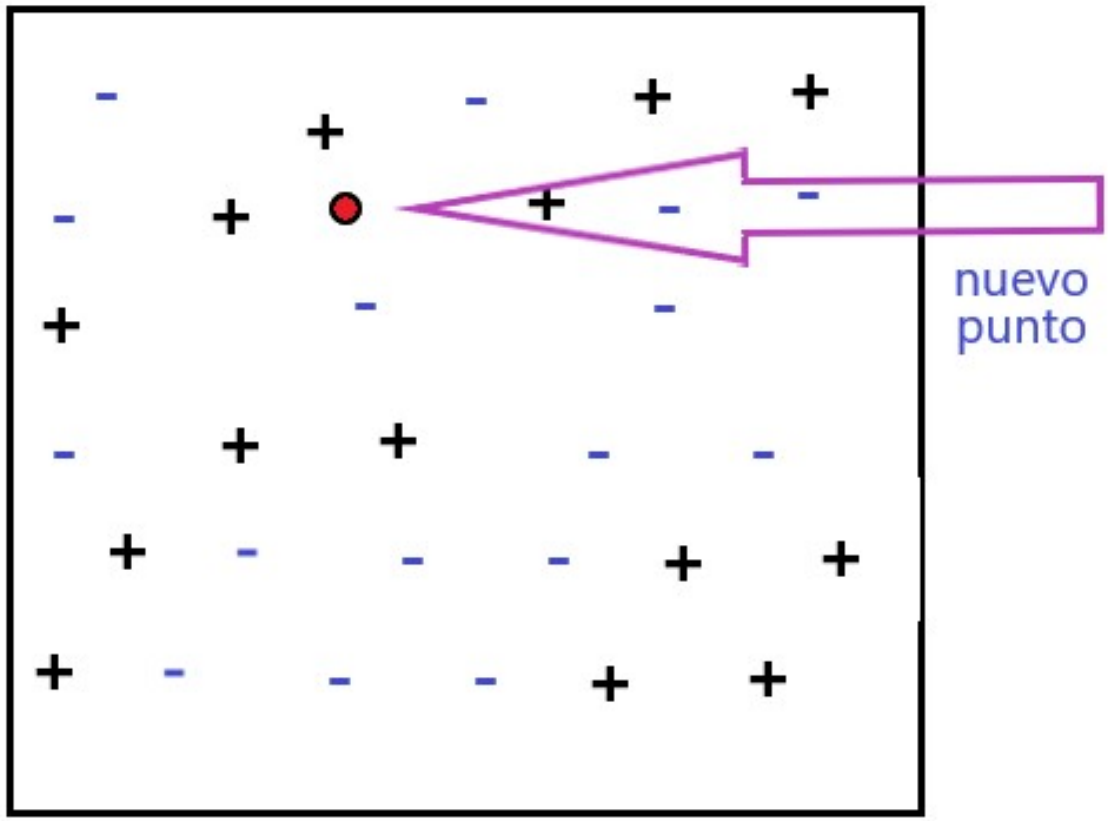
# Idea.... :)



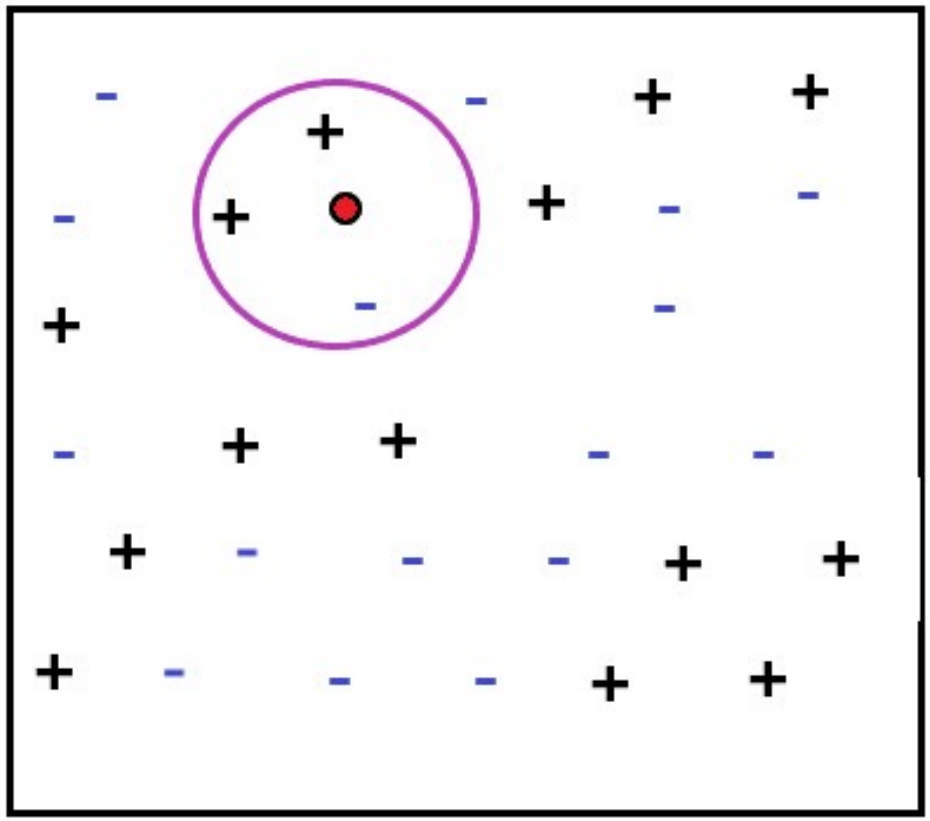
# Miremos a los Vecinos más cercanos ($k$NN: $k$-nearest neighbors)



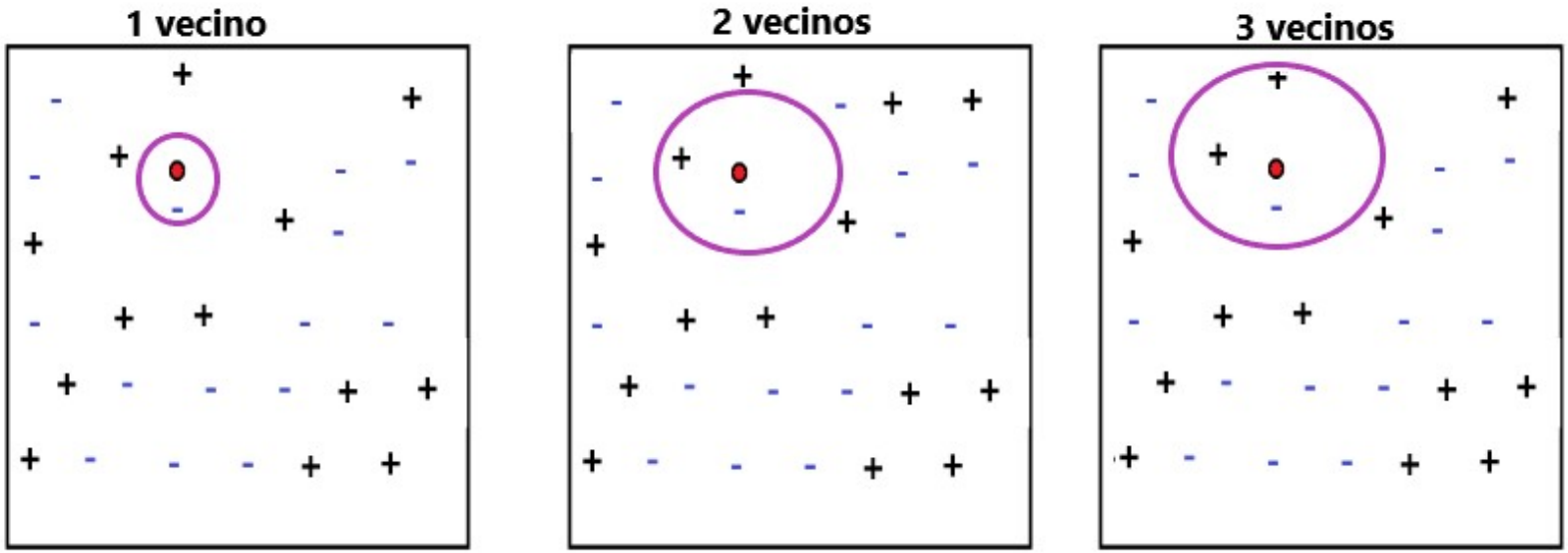
# Miremos a los Vecinos más cercanos ($k$NN: $k$-nearest neighbors)



# Miremos a los 3 Vecinos más cercanos ($k$NN: 3-nearest neighbors)



# Miremos a los Vecinos más cercanos ($k$NN: $k$-nearest neighbors)



# $k-$Vecinos más cercanos ($k$NN: $k$-nearest neighbors)

El método de $k-$Vecinos más cercanos es uno de los métodos existentes para estimar la distribución condicional de $Y$ dado $X$ y después clasificar una observación en la clase con la mayor probabilidad estimada.

- Elegimos $k$ un entero positivo y un punto $x$ para clasificar.
- Requiere una noción de distancia. Dados $\mathbf{z}, \mathbf{w} \in \mathbb{R}^q$

  Distancia euclídea: $d(\mathbf{z}, \mathbf{w}) = \sqrt{\sum_{i=1}^{q}(z_i - w_i)^2}$
- El clasificador $k$NN identifica el conjunto de los $k$ puntos más cercanos a $x$. Sea $N_x$ dicho conjunto.
- Estima a $P(Y = 1 \mid X = x)$ por la fracción de puntos en $N_x$ cuya etiqueta es igual a 1:

$$\widehat{\mathbb{P}}(Y = 1 \mid X = x) = \frac{1}{k} \sum_{i \in N_x} \mathbb{I}(y_i = 1)$$

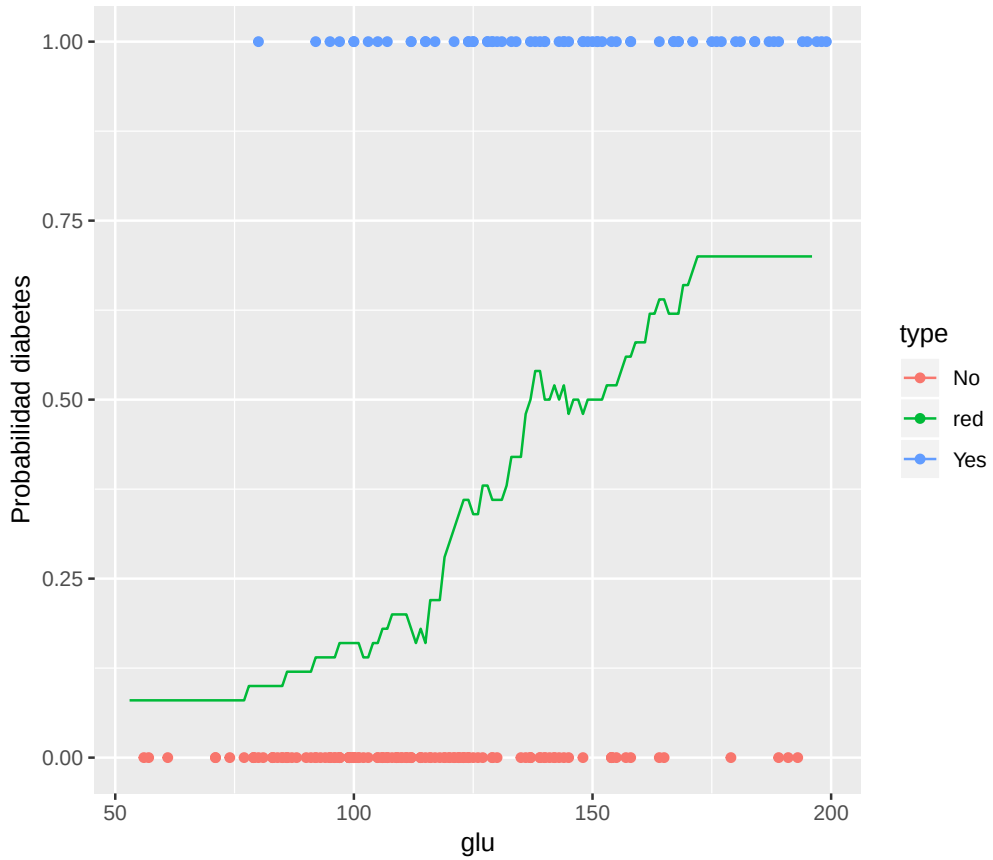
- El parámetro $k$ de este método puede elegirse por Convalizadión Cruzada.

# Ejemplo: Datos PIMA

```
> rm(list=ls())
> library(kknn)
> library(tidyr)
> library(ggplot2)
> #
> diabetes_train <- as.data.frame(MASS::Pima.tr)
> diabetes_test<- as.data.frame(MASS::Pima.te)
> #
> datos_graf <- data.frame(glu = seq(53,196, 1))
> salida <- kknn(type ~ glu, train = diabetes_train,  k = 50,
+                test = datos_graf, kernel = 'rectangular')
> #
> pepe<- as.matrix(salida$prob)
> datos_graf$probas<- pepe[,2]
```

# Ejemplo: Datos PIMA

```
> aa=ggplot(diabetes_train, aes(x = glu, y= as.numeric(type=='Yes'), colour = ty
+   geom_point() +
+   geom_line(data=datos_graf,aes(x=glu, y = probas,colour="red")) +
+   ylab('Probabilidad diabetes')+coord_fixed(ratio=150)
```



Vayamos al Ejercicico 1 de la Guía 2

# Algunas características

- Es un método muy intuitivo en el que a un nuevo punto se le asigna una categoría por voto de la mayoría entre los $k$ vecinos más cercanos.
- Se generaliza muy fácilmente a una problema con más de dos clases.
- Se pueden usar distintas distancias.
- Si $k$ es muy pequeño es muy sensible al ruido, si es muy grande podría incluir vecinos de otras clases.

# Algunas características

- Es un método muy intuitivo en el que a un nuevo punto se le asigna una categoría por voto de la mayoría entre los $k$ vecinos más cercanos.
- Se generaliza muy fácilmente a una problema con más de dos clases.
- Se pueden usar distintas distancias.
- Si $k$ es muy pequeño es muy sensible al ruido, si es muy grande podría incluir vecinos de otras clases.
- Para prevenir que alguno de los atributos tenga más influencia en la medida de distancia que otros se suele escalar, de esta manera una distancia d signifique lo mismo para el atributo 1 y para el 2, por ejemplo.
- Atributos irrelevantes podrían incrementar la distancia artificialmente a casos similares.
- Maldición de la dimensión.

# Algunas características

- Es un método muy intuitivo en el que a un nuevo punto se le asigna una categoría por voto de la mayoría entre los $k$ vecinos más cercanos.
- Se generaliza muy fácilmente a una problema con más de dos clases.
- Se pueden usar distintas distancias.
- Si $k$ es muy pequeño es muy sensible al ruido, si es muy grande podría incluir vecinos de otras clases.
- Para prevenir que alguno de los atributos tenga más influencia en la medida de distancia que otros se suele escalar, de esta manera una distancia d signifique lo mismo para el atributo 1 y para el 2, por ejemplo.
- Atributos irrelevantes podrían incrementar la distancia artificialmente a casos similares.
- Maldición de la dimensión.
- La clasificación de nuevos registros es más costosa que con otros métodos. Es un clasificador de aprendizaje perezoso (lazy).
- No construye un modelo explícito.

# Otro Enfoque: Regresión Logística

La regla de óptima de Bayes depende de:

$$p(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$$

$\mathbf{X} = (X_1, \ldots, X_p)^t$ es un vector de $p$ covariables.

¿Y si modelamos esta probabilidad en función de $\mathbf{X}$?

# Otro Enfoque: Regresión Logística

La regla de óptima de Bayes depende de:

$$p(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$$

$\mathbf{X} = (X_1, \ldots, X_p)^t$ es un vector de $p$ covariables.

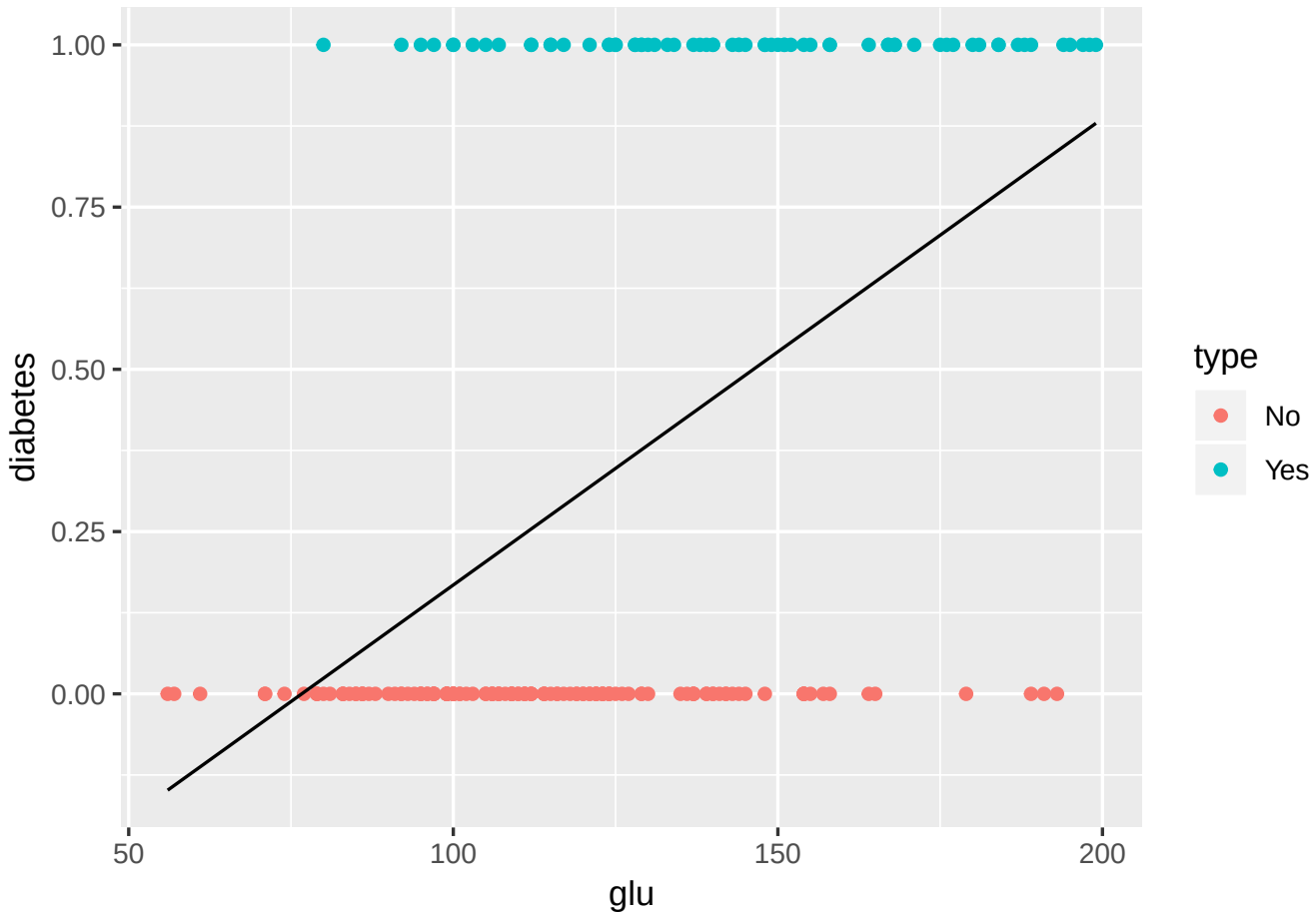
¿Y si modelamos esta probabilidad en función de $\mathbf{X}$?

Tenemos que tener algunos cuidados...

# ¿Qué pasa si proponemos un modelo lineal?

```
> datos_ent <- as.data.frame(MASS::Pima.tr)
> attach(datos_ent)
> diabetes<- 1*(type=="Yes")
> datos_ent$diabetes<- diabetes
> mean(diabetes)

[1] 0.34

> valor_new <- 70
> m_lineal <- lm(diabetes ~ glu, data = datos_ent)
> predict(m_lineal, data.frame(glu = valor_new))

          1
-0.04782927

>
```

# ¿Qué pasa si proponemos un modelo lineal?



# Regresión Logística

Pensemos que tenemos una sola variable $X$

Notemos que

$$0 \quad \leq \quad p(x) \leq 1$$

# Regresión Logística

Pensemos que tenemos una sola variable $X$

Notemos que

$$0 \quad \leq \quad p(x) \leq 1$$

Sin embargo, los **odds** (o chances en castellano...)

$$\frac{p(x)}{1-p(x)} \geq 0$$

# Regresión Logística

Pensemos que tenemos una sola variable $X$

Notemos que

$$0 \quad \leq \quad p(x) \leq 1$$

Sin embargo, los **odds** (o chances en castellano...)

$$\frac{p(x)}{1-p(x)} \geq 0$$

Tomando logaritmo:

$$-\infty \quad < \quad \log\left(\frac{p(x)}{1-p(x)}\right) < \infty$$

Podríamos modelar:

$$\log\left(\frac{p(x)}{1-p(x)}\right) \quad = \quad \beta_0 + \beta_1\, x$$

# Regresión Logística

Haciendo el camino inverso:

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

$$p(x) = p(x,\beta) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

# Regresión Logística

Haciendo el camino inverso:

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

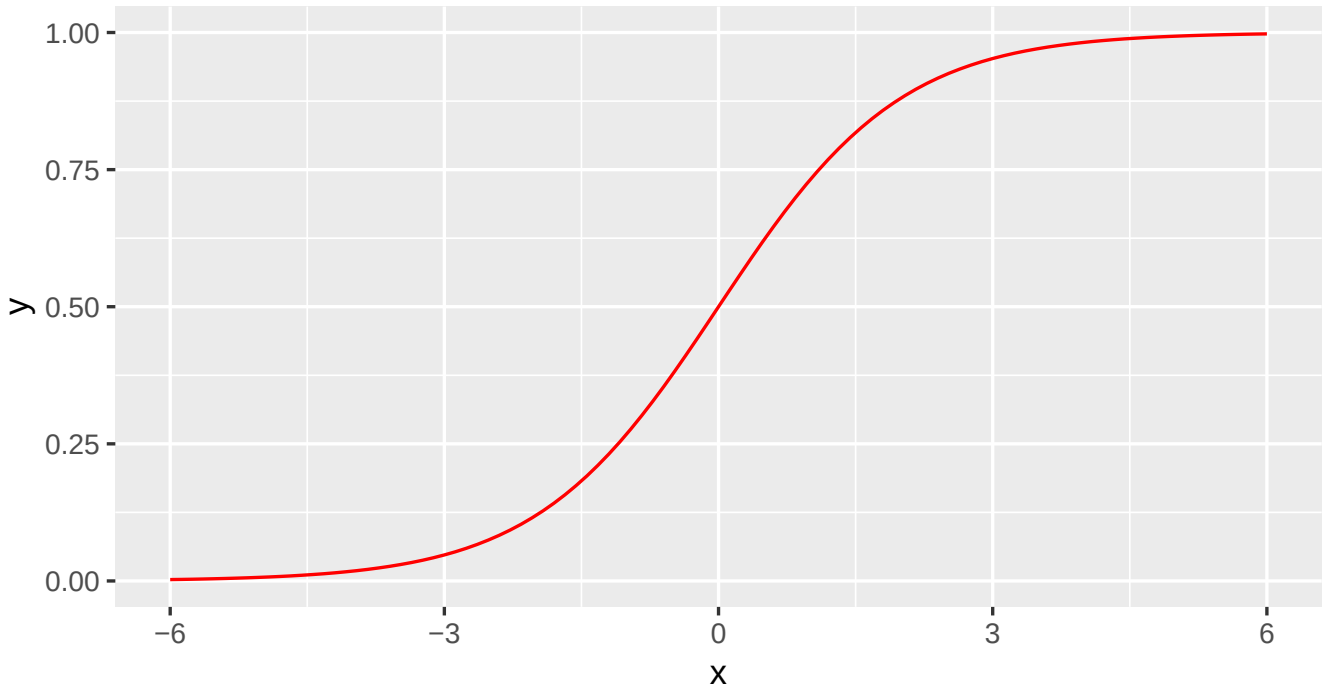
$$p(x) = p(x,\beta) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$$

# Función logística

La función logística está definida por

$$p(x) = \frac{e^x}{1 + e^x}$$

que corresponde en nuestro caso a tomar $\beta_0 = 0$ y $\beta_1 = 1$.



# ¿Cómo se interpreta aquí al coeficiente $\beta_1$?

$$odds(x) = \frac{p(x)}{1-p(x)} \quad = \quad e^{\beta_0+\beta_1\, x}$$

Si aumentamos a $x$ en una unidad, tenemos que

$$odds(x+1) = \frac{p(x+1)}{1-p(x+1)} \quad = \quad e^{\beta_0+\beta_1\, (x+1)}$$

por lo tanto

$$\frac{odds(x+1)}{odds(x)} = e^{\beta_1} \Rightarrow \log\left(\frac{odds(x+1)}{odds(x)}\right) = \beta_1$$

# Regresión Logística

Con una sola variable

$$p(x) = p(x, \beta) \quad = \quad \frac{e^{\beta_0 + \beta_1 \, x}}{1 + e^{\beta_0 + \beta_1 \, x}}$$

# Regresión Logística

Con una sola variable

$$p(x) = p(x, \beta) \quad = \quad \frac{e^{\beta_0 + \beta_1 \, x}}{1 + e^{\beta_0 + \beta_1 \, x}}$$

En general, tendremos para un vector de covariables

Modelo de regresión logística:

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \tag{1}$$

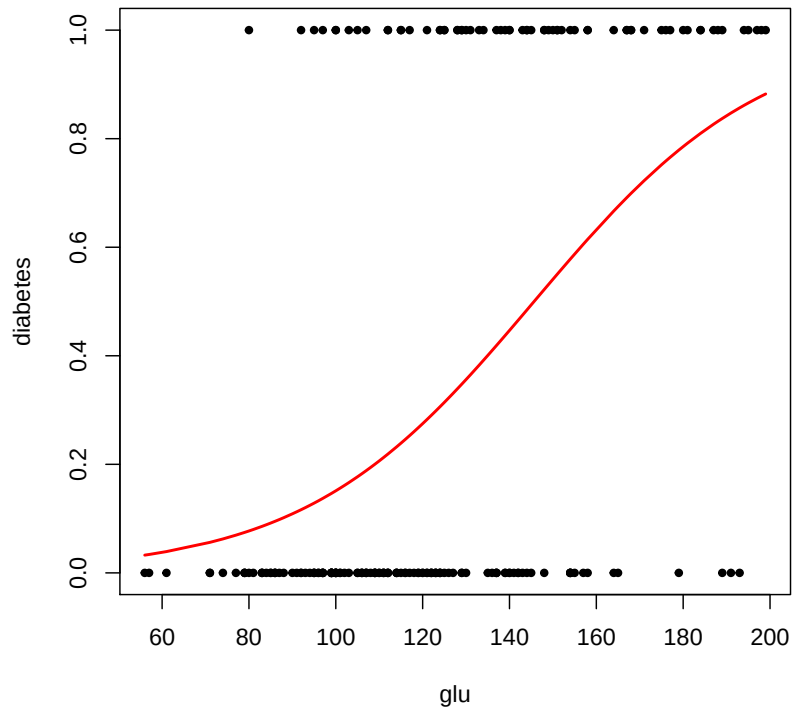
**Recordemos qué son los estimadores de máxima verosimilitud....**

# Volvamos a los datos PIMA

```
> salida_glm <- glm(diabetes ~ glu, family=binomial(),data = datos_ent)
> summary(salida_glm)

Call:
glm(formula = diabetes ~ glu, family = binomial(), data = datos_ent)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9714  -0.7795  -0.5292   0.8491   2.2633

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.503636   0.836077  -6.583 4.62e-11 ***
glu          0.037784   0.006278   6.019 1.76e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 207.37  on 198  degrees of freedom
AIC: 211.37

Number of Fisher Scoring iterations: 4
```

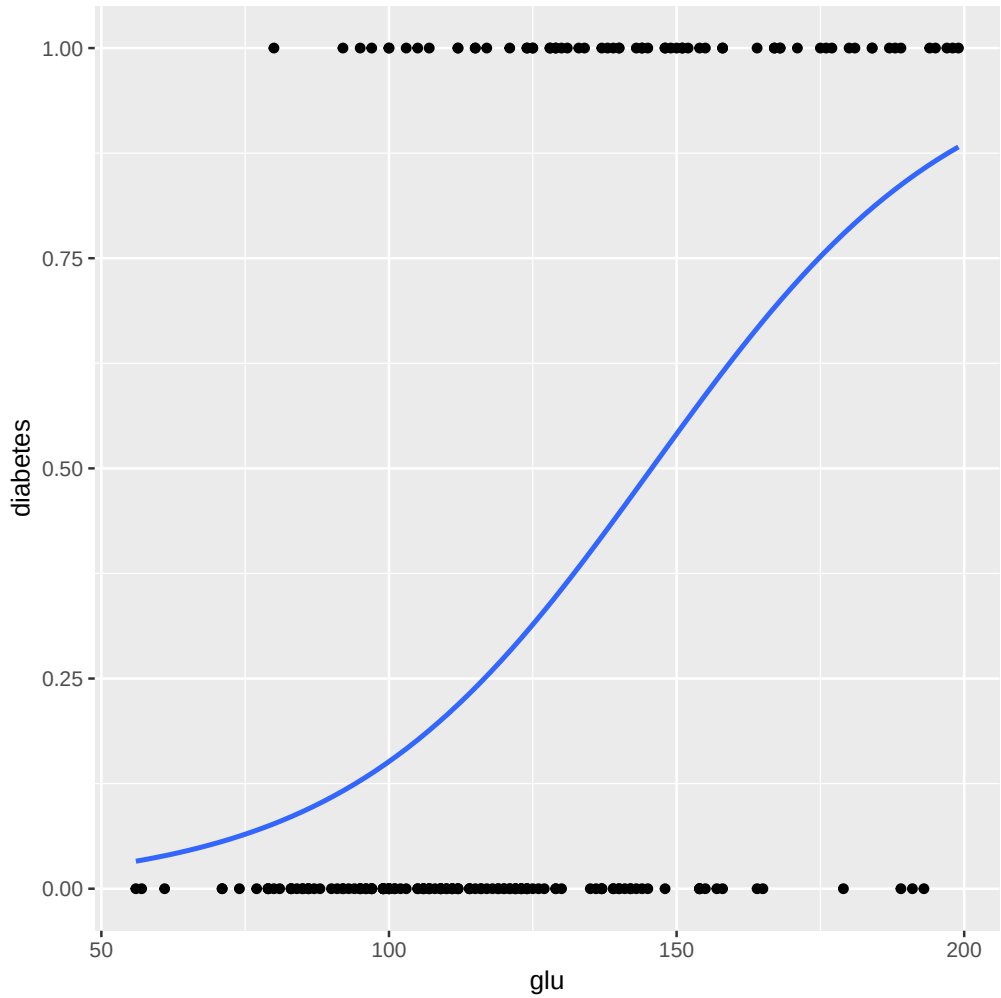
# Grafiquemos

```
> orden<- order(glu)
> plot(glu[orden],diabetes[orden],pch=20,xlab="glu",ylab="diabetes")
> lines(glu[orden],salida_glm$fitted.values[orden],col="red",lwd=2)
```



# Grafiquemos con ggplot2

```
> library(ggplot2)
> aa<-ggplot(datos_ent, aes(x=glu, y=diabetes)) + geom_point() +
+   stat_smooth(method="glm", method.args=list(family="binomial"), se=F)
```



# Algunas Generalidades sobre Inferencia Estadística

## Recordemos.....

- Ingredientes: datos generados por un mecanismo aleatorio: por ej., tiramos una moneda al aires sucesivas veces.

- Objetivo: inferir *algo relacionado* con el mecanismo (aleatorio) que genera los datos:

  por ej., ¿cuál es la probabilidad de obtener cara con nuestra moneda?

- Mecanismo: Función de distribución.
  - **Caso discreto**: función de probabilidad puntual (Binomial, Poisson....)
  - **Caso continuo**: función de densidad (Normal, t, Gamma....)

- Modus Operandi: hacer *alguna cuenta* con los datos para obtener un valor que *se parezca* al que queremos inferir.

# Muestra

- Muestra aleatoria: $Y_1, \cdots, Y_n$ variables aleatorias i.i.d. (independientes e idénticamente distribuidas)

- Datos u observaciones: $\mathbf{y} = y_1, \ldots, y_n$ constituyen una realización de la muestra aleatoria.

# Inferencia Estadística

- Datos: $(Y_i)_{i\geq 1}$ i.i.d. $Y_i \sim F$, $F \in \mathcal{F}$ familia de distribuciones posibles para nuestro problema

- Objetivo: inferir *algo relacionado* con el mecanismo que genera los datos:
  - $\mathbb{E}_F[Y]$
  - $\mathbb{V}_F(Y)$
  - $\mathbb{P}_F(Y > 0{,}4)$
  - $F$.

# Modelos Paramétricos: $M$

Asumimos que la función de distribución que genera los datos pertenece a una familia

$$M = \{F(\cdot, \theta) \, , \theta \in \Theta\} \, ,$$

siendo $\Theta \subset \mathbb{R}^k$, para algún $k$.

- Caso discreto: $M = \{p(\cdot, \theta) \, , \theta \in \Theta\} \, ,$
- Caso continuo: $M = \{f(\cdot, \theta) \, , \theta \in \Theta\} \, .$

# Modelos Paramétricos

## Ejemplos discretos

$$M = \{p(\cdot, \theta)\ , \theta \in \Theta\}\ ,$$

- Bernoulli: $Y \sim B(1, \theta)$:

$$p(y, \theta) = \theta^y (1-\theta)^{1-y}, \quad y = 0, 1$$

$p \in \Theta = [0, 1] \subset \mathbb{R}$.

- Poisson: $Y \sim \mathbb{P}(\lambda)$:

$$p(y, \lambda) = e^{-\lambda} \frac{\lambda^y}{y!}, \quad \lambda \in \Theta = \mathbb{R}_{>0}$$

# Modelos Paramétricos

## Ejemplos continuos

$$M = \{f(\cdot, \theta)\ , \theta \in \Theta\}\ .$$

- Normal: $Y \sim N(\mu, \sigma^2)$:

$$f(y, \theta) = (2\pi\sigma^2)^{-1/2} e^{-(y-\mu)^2/2\sigma^2}, \quad \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$$

- Exponencial: $Y \sim E(\lambda)$:

$$f(y, \lambda) = \lambda e^{-\lambda y} \mathbb{I}_{[0,\infty)}(y), \quad \lambda \in \Theta = \mathbb{R}_{>0}$$

# ¿Cómo estimamos los parámetros desconocidos?

## Idea: Verosimilitud

- Antes de realizar el *experimento* el **resultado** es **desconocido**.
- Cuando los parámetros son **conocidos**, las probabilidades nos permiten predecir un resultado **desconocido**, por ejemplo

$$P(\textit{resultado}|\theta) \overset{\text{por ej. caso Binomial}}{=} P(y|n,p) = \binom{n}{y} p^y (1-p)^{(n-y)}$$

# ¿Cómo estimamos los parámetros desconocidos?

## Idea: Verosimilitud

- Antes de realizar el *experimento* el **resultado** es **desconocido**.
- Cuando los parámetros son **conocidos**, las probabilidades nos permiten predecir un resultado **desconocido**, por ejemplo

$$P(\textit{resultado}|\theta) \overset{\text{por ej. caso Binomial}}{=} P(y|n,p) = \binom{n}{y} p^y (1-p)^{(n-y)}$$

Ahora se invierte el paradigma:

- Al realizar el experimento el **resultado** se hace **conocido**: **dato**.
- Nos interesa conocer cuán **verosímil** es que un determinado parámetro haya generado el dato.

# Verosimilitud

## Ejemplo

- Monedas, cara= 1, ceca= 0.
- Moneda bolsillo derecho: equilibrada
- Moneda bolsillo izquierdo: probabilidad de cara es 0,8.
- Objetivo: identificar la moneda a partir de una muestra. En $n = 100$ lanzamientos se observa la muestra

$$\mathbf{y} = \underbrace{1,\ldots,1}_{\textbf{12veces}}\underbrace{0,\ldots,0}_{\textbf{5veces}}\underbrace{1,\ldots,1}_{\textbf{23veces}}\underbrace{0,\ldots,0}_{\textbf{8veces}}\underbrace{1,\ldots,1}_{\textbf{15veces}}\underbrace{0,\ldots,0}_{\textbf{3veces}}$$
$$\underbrace{1,\ldots,1}_{11veces}\underbrace{0,\ldots,0}_{4veces}\underbrace{1,\ldots,1}_{13veces}\underbrace{0,\ldots,0}_{6veces}$$

- ¿Cuál de las dos monedas diría que está utilizando?

# Función de Verosimilitud

- $L(0{,}8; \mathbf{y})$ = probabilidad de observar la muestra $\mathbf{y}$ con moneda de $p = 0{,}8$.
- $L(0{,}5; \mathbf{y})$ = probabilidad de observar la muestra $\mathbf{y}$ con moneda de $p = 0{,}5$ (equilibrada).

# Función de Verosimilitud

- $L(0{,}8; \mathbf{y})$ = probabilidad de observar la muestra $\mathbf{y}$ con moneda de $p = 0{,}8$.

$$L(0{,}8; \mathbf{y}) = \underbrace{0{,}8 \ldots 0{,}8}_{12 veces} \underbrace{0{,}2 \ldots 0{,}2}_{5 veces} \underbrace{0{,}8 \ldots 0{,}8}_{23 veces} \underbrace{0{,}2 \ldots 0{,}2}_{8 veces} \underbrace{0{,}8 \ldots 0{,}8}_{15 veces} \underbrace{0{,}22 \ldots 0{,}2}_{3 veces}$$

$$\underbrace{0{,}8 \ldots 0{,}8}_{11 veces} \underbrace{0{,}2 \ldots 0{,}2}_{4 veces} \underbrace{0{,}8 \ldots 0{,}8}_{13 veces} \underbrace{0{,}2 \ldots 0{,}2}_{6 veces} =$$

$$= (0{,}8)^{74}(0{,}2)^{26} = 4{,}523128\ 10^{-26}$$

siendo 74 el número de caras observadas en las $n = 100$ repeticiones.

# ¿Y ahora?

- $\mathbf{y}$ con 74 caras, 26 cecas.
- $L(0{,}8; \mathbf{y}) = 0{,}8^{74}0{,}2^{26} = 4{,}523128\ 10^{-26}$
- $L(0{,}5; \mathbf{y}) = (1/2)^{74}(1/2)^{26} = 7{,}888609\ 10^{-31}$
- ¿cuál de las dos monedas diríamos que se está utilizando?

# Propuesta de Máxima Verosimilitud

La propuesta de máxima verosimilitud consiste en pensar que la moneda que estamos utilizando es aquella para la cual los valores observados resultan más probables. Es decir, elegimos la moneda que maximiza la probabilidad de los valores observados. Siendo que

$$L(0{,}8; \mathbf{y}) > L(0{,}5; \mathbf{y})$$

concluimos que se está utilizando la moneda no equilibrada.

# Sintetizando

- $Y_i \sim B(1, \theta)$, $\theta \in [0, 1]$.
- 74 caras en $n = 100$ repeticiones.
- ¿$\theta$?
- vamos a elegir aquel valor del parámetro para el cuál los valores observados tienen mayor probabilidad de ocurrir.

# Función de Verosimilitud

- $L(\theta; \mathbf{y})$: Mide cuál es la probabilidad de observar nuestra realización $\mathbf{y}$ cuando la probabilidad de cara es $\theta$.
- en $\mathbf{y}$ de tamaño 100 hay 74 caras

$$L(\theta; \mathbf{y}) = \theta^{74}(1-\theta)^{26}$$

- queremos maximizar $L(\theta)$.
- $\Longleftrightarrow$ maximizar $l(\theta; \mathbf{y}) = \ln(\mathbf{L}(\theta; \mathbf{y}))$
- $l(\theta; \mathbf{x}) = 74 \ln(\theta; \mathbf{y}) + 26 \ln(1-\theta; \mathbf{y})$, se maximiza en $74/100$.
- Tenemos así que el valor estimado del parámetro con la muestra dada es $= 0{,}74$.

# Función de verosimilitud basada en **y**

## Caso Binomial en general

- $L(\theta; \mathbf{y})$: probabilidad de observar **y** cuando la probabilidad de cara es $\theta$.

- $(Y_i)_{i \geq 1}$ i.i.d., $Y_i \sim B(1, \theta)$, $p(y, \theta) = \theta^y (1-\theta)^{1-y}$

$$
\begin{aligned}
L(\theta; \mathbf{y}) &= \prod_{i=1}^{n} p(y_i, \theta) \\
&= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n - \sum_{i=1}^n y_i}
\end{aligned}
$$

# Función de verosimilitud basada en **y**

## Caso Binomial en general

- $L(\theta; \mathbf{y})$: probabilidad de observar **y** cuando la probabilidad de cara es $\theta$.

- $(Y_i)_{i \geq 1}$ i.i.d., $Y_i \sim B(1, \theta)$, $p(y, \theta) = \theta^y (1-\theta)^{1-y}$

$$\begin{aligned} L(\theta; \mathbf{y}) &= \prod_{i=1}^{n} p(y_i, \theta) \\ &= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n - \sum_{i=1}^n y_i} \end{aligned}$$

- Maximizar $L(\theta; \mathbf{y})$ equivale a maximizar $\ell(\theta; \mathbf{y}) = \log \mathbf{L}(\theta; \mathbf{y})$ ya que log es estrictamente creciente.

# Función de verosimilitud basada en **y**

## Caso Binomial en general

- Maximizar $L(\theta; \mathbf{y})$ equivale a maximizar $\ell(\theta; \mathbf{y}) = \log \mathbf{L}(\theta; \mathbf{y})$

$$\ell(\theta; \mathbf{y}) \quad = \quad \sum_{i=1}^{n} y_i \ \log(\theta) + (n - \sum_{i=1}^{n} y_i) \ \log(1-\theta)$$

- Derivemos e igualemos a 0:

$$\frac{\partial \ell(\theta; \mathbf{y})}{\partial \theta} \quad = \quad \frac{\sum_{i=1}^{n} y_i}{\theta} + \frac{n - \sum_{i=1}^{n} y_i}{1-\theta} = 0$$

- El punto crítico resulta $\bar{y}_n$
- Como cada vez que se observan los datos $y_1, \ldots, y_n$ $\bar{x}_n$ es el punto crítico

$$\Rightarrow \widehat{\theta} = \bar{Y}_n$$

# EMV: caso binomial

## Revisemos lo que hicimos

- $\mathbf{y} = \mathbf{y_1}, \ldots, \mathbf{y_n}$.
- $L(\theta; \mathbf{y}) = \theta^{\sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{y_i}} (\mathbf{1} - \theta)^{\mathbf{n} - \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{y_i}}$
- Maximizar $L(\theta; \mathbf{y})$
- Maximizar $\ell(\theta; \mathbf{y}) = \log(\mathbf{L}(\theta; \mathbf{y}))$:
- $\ell(\theta; \mathbf{y})$ se maximiza en $\frac{1}{n} \sum_{i=1}^{n} y_i$.
- EMV: $\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$

# EMV

## Caso discreto

- Modelo: $M = \{p(\cdot, \theta)\ , \theta \in \Theta\}$.
- $\mathbf{y}= y_1, \cdots, y_n$ realización de $Y_1, \cdots, Y_n$ i.i.d.
- Función de verosimilitud asociada a $\mathbf{y}= y_1, \cdots, y_n$:

$$L(\cdot\ ;\ \mathbf{y}\ )\ : \Theta \to \mathbb{R}$$

$$L(\theta\ ;\ \mathbf{y}\ ) = \mathbb{P}(Y_1 = y_1, \cdots, Y_n = y_n)\ , Y_i \sim p(\cdot, \theta)\ .$$

$$L(\theta\ ;\ \mathbf{y}\ ) = \prod_{i=1}^{n} p(y_i, \theta)\ ,$$

- Propuesta de máxima verosimilitud:

$$h_n(\mathbf{y}) = \underset{\theta \in \Theta}{\text{argmax}}\ L(\theta\ ;\ \mathbf{y}\ )\ .$$

o sea

$$L(h_n(\mathbf{y}), \mathbf{y}) \geq L(\theta, \mathbf{y})$$

# Volvamos a nuestro caso: Regresión Logística

Estamos ante un problema un poco más complejo porque aquí las probabilidades están relacionadas a través de una función link con el parámetro $\beta$.

El EMV de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ se halla maximizando

$$L(\mathbf{b}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{b})^{y_i} (1 - p(\mathbf{x}_i, \mathbf{b}))^{1-y_i}$$

La log verosimilitud resulta

$$\ell(\mathbf{b}) = \log L(\mathbf{b}) = \sum_{i=1}^{n} y_i \log p(\mathbf{x}_i, \mathbf{b}) + (1 - y_i) \log(1 - p(\mathbf{x}_i, \mathbf{b}))$$

# Volvamos a nuestro caso: Regresión Logística

Estamos ante un problema un poco más complejo porque aquí las probabilidades están relacionadas a través de una función link con el parámetro $\beta$.

El EMV de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ se halla maximizando

$$L(\mathbf{b}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{b})^{y_i} (1 - p(\mathbf{x}_i, \mathbf{b}))^{1-y_i}$$

La log verosimilitud resulta

$$\ell(\mathbf{b}) = \log L(\mathbf{b}) = \sum_{i=1}^{n} y_i \log p(\mathbf{x}_i, \mathbf{b}) + (1 - y_i) \log(1 - p(\mathbf{x}_i, \mathbf{b}))$$

Típicamente, para hallar el EMV se deriva la log–verosimilitud $\ell(\boldsymbol{\beta})$ e iguala a 0:

$$\frac{\partial \ell}{\partial b_j} = \sum_{i=1}^{n} (y_i - p(\mathbf{x}_i, \mathbf{b})) x_{ij} = 0 \quad j = 0, 1 \ldots, p$$

# Volvamos a nuestro caso: Regresión Logística

$$\frac{\partial \ell}{\partial b_j} = \sum_{i=1}^{n} (y_i - p(\mathbf{x}_i, \mathbf{b}))x_{ij} = 0 \quad j = 0, 1 \ldots, p$$

es de la forma

$$\mathbf{X}^t(\mathbf{Y} - \mathbf{P}(\mathbf{b}))$$

donde $\mathbf{Y}$ y $\mathbf{P}$ representan a los vectores con componentes $y_i$ y $p(\mathbf{x}_i, \mathbf{b})$.

El estimador de máxima versosimilitud de $\beta$, $\widehat{\beta}$, resuelve

$$\mathbf{X}^t(\mathbf{Y} - \mathbf{P}(\widehat{\beta})) = \mathbf{0}$$

# En efecto

Notemos que bajo el modelo logístico $p(\mathbf{x}, \boldsymbol{\beta}) = h(\mathbf{x}^t \boldsymbol{\beta})$, siendo $h(t) = \dfrac{1}{1 + e^{-t}}$

Es fácil ver que

$$h'(t) = h(t)(1 - h(t))$$

Cada término es de la forma:

$$y \log p(\mathbf{x}, \mathbf{b}) + (1 - y) \log(1 - p(\mathbf{x}, \mathbf{b}))$$

por lo tanto, su derivada parcial respecto de $b_j$ es:

$$\frac{y}{p(\mathbf{x}, \mathbf{b})} p(\mathbf{x}, \mathbf{b})(1 - p(\mathbf{x}, \mathbf{b}))x_j - \frac{(1 - y)}{1 - p(\mathbf{x}, \mathbf{b})} p(\mathbf{x}, \mathbf{b})(1 - p(\mathbf{x}, \mathbf{b}))x_j$$

que resulta

$$(y - p(\mathbf{x}, \mathbf{b}))\, x_j$$

# Método de de Newton–Raphson

Supongamos que queremos resolver

$$\mathbf{f}(\mathbf{z}) = \mathbf{f}(z_1, \ldots, z_q) = \begin{pmatrix} f_1(z_1, \ldots, z_q) \\ \vdots \\ f_q(z_1, \ldots, z_q) \end{pmatrix} = 0\,.$$

Supongamos además que $\xi$ es solución y que $\mathbf{z}_0$ es un punto próximo a $\xi$. Usando una expansión de Taylor de primer orden alrededor de $\mathbf{z}_0$ tenemos que

$$0 = \mathbf{f}(\xi) \approx \mathbf{f}(\mathbf{z}_0) + \nabla\mathbf{f}(\mathbf{z}_0)(\xi - \mathbf{z}_0)$$

donde

$$\nabla\mathbf{f}(\mathbf{z}_0) = \begin{pmatrix} \frac{\partial f_1}{\partial z_1} & \cdots & \frac{\partial f_1}{\partial z_q} \\ \cdots & \ddots & \cdots \\ \frac{\partial f_q}{\partial z_1} & \cdots & \frac{\partial f_q}{\partial z_q} \end{pmatrix}_{\mathbf{z}=\mathbf{z}_0}\,.$$

# Método de de Newton–Raphson

Luego,

$$\xi = \mathbf{z}_0 - [\nabla \mathbf{f}(\mathbf{z}_0)]^{-1}\mathbf{f}(\mathbf{z}_0)$$

El método de Newton Raphson es un método iterativo con un punto inicial $\mathbf{z}_0$ y tal que en el paso $k+1$ se actualiza el valor buscado de la siguiente forma

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - [\nabla \mathbf{f}(\mathbf{z}^{(k)})]^{-1}\mathbf{f}(\mathbf{z}^{(k)})$$

Para el caso que nos interesa resolver resultaría

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left[\ell''(\boldsymbol{\beta}^{(k)})\right]^{-1} \ell'(\boldsymbol{\beta}^{(k)})$$

Si $\ell(\boldsymbol{\beta})$ fuera cuadrática, entonces $\ell'(\boldsymbol{\beta})$ sería lineal y el algoritmo iterativo convergería en un sólo paso a partir de un punto inicial.

# Método de de Newton–Raphson

En problemas regulares, el log–likelihood se hace aproximadamente cuadrático a medida que $n$ crece.

En estas situaciones el método de NR funcionará bien, mientras que en muestras pequeñas y con log–likelihoods alejados de una cuadrática NR podría no converger.

# Caso logístico

Si usamos las relaciones anteriores resulta que las derivadas segundas:

$$\frac{\partial^2 \ell(\mathbf{b})}{\partial b_j \partial b_k} = -\sum_{i=1}^{n} p(\mathbf{x}_i, \mathbf{b})\,(1 - p(\mathbf{x}_i, \mathbf{b}))x_{ij}x_{ik} = -\mathbf{X}^t\mathbf{W}\mathbf{X}$$

donde $\mathbf{W} = diag(p(\mathbf{x}_1, \mathbf{b})(1 - p(\mathbf{x}_1, \mathbf{b})), \ldots, p(\mathbf{x}_n, \mathbf{b})(1 - p(\mathbf{x}_n, \mathbf{b})))$

$\mathbf{W}$ depende de los parámetros que queremos estimar: en el paso $t$ habrá que evaluarla en $\beta^{(t)}$.

Newton–Raphson resulta:

$$\beta^{(k+1)} = \beta^{(k)} + \left(\mathbf{X}^t\mathbf{W}^{(k)}\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\mathbf{Y} - \mathbf{P}^{(k)}\right).$$

$\mathbf{Y}$ y $\mathbf{P}^{(k)}$: vectores con componentes $y_i$ y $p(\mathbf{x}_i, \beta^{(k)})$, respectivamente.

# Caso logístico

Es interesante ver que si en cada paso $k$ creamos la pseudo-observación

$$z_i^{(k)} = \mathbf{x}_i^t \beta^{(k)} + \frac{y_i - p(\mathbf{x}_i, \beta^{(k)})}{p(\mathbf{x}_i, \beta^{(k)}) i (1 - p(\mathbf{x}_i, \beta^{(k)}))} = \mathbf{x}_i^t \beta^{(k)} + \frac{y_i - \pi_i^{(k)}}{\pi_i^{(k)}(1 - \pi_i^{(k)})}$$

entonces, si $\mathbf{z}$ es el vector formado con estas pseudo-observaciones

$$\beta^{(k+1)} = \left(\mathbf{X}^t \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{W}^{(k)} \mathbf{z}^{(k)}$$

donde
$\mathbf{W}^{(k)} = diag(p(\mathbf{x}_1, \beta^{(k)})(1 - p(\mathbf{x}_1, \beta^{(k)})), \ldots, p(\mathbf{x}_n, \beta^{(k)})(1 - p(\mathbf{x}_n, \beta^{(k)})))$

# EMV: Regresión Logística

Típicamente, para hallar el EMV se deriva la log–verosimilitud $\ell(\mathbf{b})$ e iguala a 0:

$$\frac{\partial \ell}{\partial b_j} = \sum_{i=1}^{n}(y_i - p(\mathbf{x}_i, \mathbf{b})\mathbf{x}_{ij} = 0 \quad j = 0, 1 \ldots, p$$

La solución de este sistema no tiene una expresión analítica y debe ser resuelto numéricamente: resolviendo numéricamente la ecuación por el método de Newton-Raphson o Fisher-scoring.