Ciencias de Datos con R: Fundamentos Estadísticos

Ana M. Bianco, Jemina García y Mariela Sued

Clases 26 - Revisitando Clasificación...

Volvamos al problema del Patito Feo

Volvamos al problema del Patito Feo

En un bosque de talas...

- Dos posibles hospedadores:

$$Y = \begin{cases} 1 & \text{rechazador} \\ 0 & \text{aceptador .} \end{cases}$$

- Se colocan $n = 8$ huevos parasitarios en un nideo elegido al azar.
- $X =$ número de huevos removidos.
- Si se remueven 5 huevos; ¿de qué clase de nido diría que se trata?
- Si se remueven 3 huevos; ¿de qué clase de nido diría que se trata?

Clasificación: Marco Teórico

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

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

- Error de Clasificación Medio (verdadero - poblacional) del clasificador g

$$\mathbb{P}(g(X) \neq Y)$$

Clasificación: Marco Teórico

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

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

- Error de Clasificación Medio (verdadero - poblacional) del clasificador g

$$\mathbb{P}(g(X) \neq Y)$$

- ...buen momento para hacer un alto...

¿Consideramos un tipo de error medio diferente?

Error de Clasificación Medio de un clasificador g

$$\mathbb{P}(g(X) \neq Y)$$

Como $g(X)$ e Y sólo toman dos valores, 0 y 1, es fácil comprobar que

$$E((Y - g(X))^2) = \mathbb{P}(g(X) \neq Y) = L(g)$$

es decir, seguimos hablando de pérdida cuadrática.

Clasificación: Marco Teórico

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

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

- Error de Clasificación Medio (verdadero - poblacional) del clasificador g

$$L(g)$$

- Objetivo (teórico): Encontrar g que minimice el error medio de clasificación

$$g^{op}$$

g^{op} Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y=1 \mid X=x) \geq \mathbb{P}(Y=0 \mid X=x) \\ 0 & \text{si} \quad \mathbb{P}(Y=0 \mid X=x) > \mathbb{P}(Y=1 \mid X=x) \end{cases}$$

Teorema: $L(g^{op}) \leq L(g)$, para todo g.

Todo muy lindo, pero.... ¿cómo implementamos la regla de clasificación?

(X, Y) vector aleatorio

Observamos los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Estimación g^{op} Optimo: Modelos discriminativos

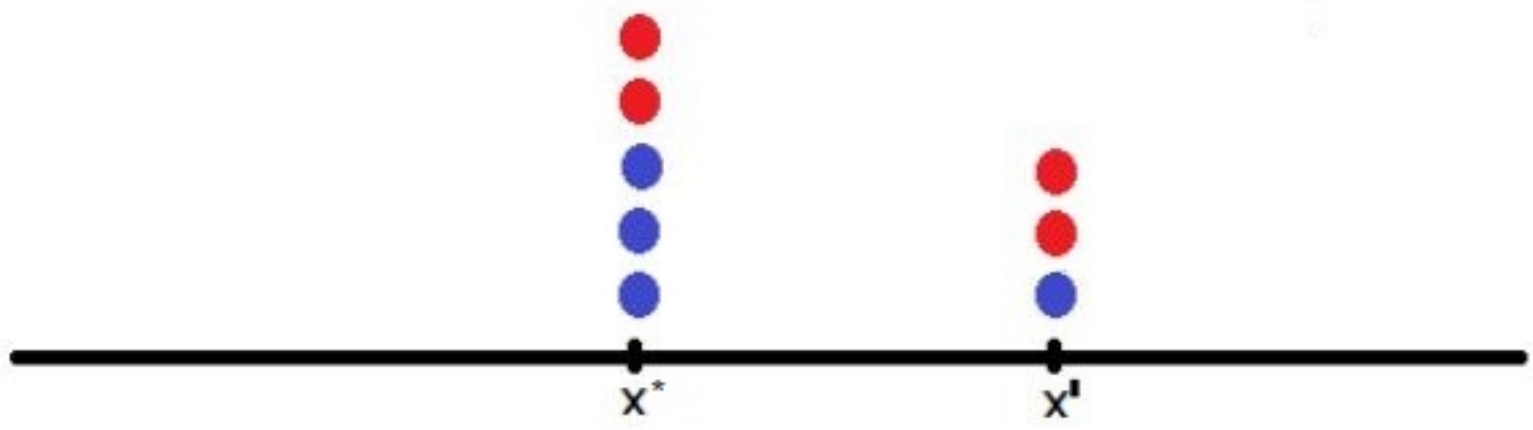
$$g^{op}(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y = 1 \mid X = x) \geq 1/2 \\ 0 & \text{si} \quad \text{c. c.} \end{cases}$$

Aquí se estima

- $\mathbb{P}(Y = 1 \mid X = x)$.

Un método democrático: regla de la mayoría

X discreta



$$\hat{\mathbb{P}}(Y = 1 \mid X = x) = \frac{\sum_{i=1}^{n} y_i I_{\{x_i = x\}}}{\sum_{i=1}^{n} I_{\{x_i = x\}}}$$

Volvamos al problema de las Alturas

Objetivo: Clasificar el género de un individuo en F o M conociendo su altura.

Ahora la variable predictora X es continua.

$k-$Vecinos más cercanos (kNN: k-nearest neighbours)

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

- Elegimos k un entero positivo y un punto x para clasificar.
- El clasificador kNN 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:

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

- Análogamente estimamos $P(Y = 0 \mid X = x)$
- $\widehat{g}_k$
- $k =$?

Otra forma: estimación por núcleos

Otra manera de estimar a $P(Y = 1 \mid X = x)$ podría ser considerar un entorno $(x - h, x + h)$ y repetir el procedimiento anterior.

- Elegimos $h > 0$ y un punto x para clasificar.
- El clasificador identifica en el intervalo $(x - h, x + h)$ los puntos con etiqueta 1 y 0
- Estima a $P(Y = 1 \mid X = x)$ por la fracción de puntos en $(x - h, x + h)$ cuya etiqueta es igual a 1:

$$\widehat{\mathbb{P}}(Y = 1 \mid X = x) = \frac{\sum_{i=1}^{n} y_i \; I_{(x-h,x+h)}(x_i)}{\sum_{i=1}^{n} I_{(x-h,x+h)}(x_i)}$$

- $\widehat{g}_h$
- $h =$?

Otra forma: estimación por núcleos

Otra manera de estimar a $P(Y = 1 \mid X = x)$ podría ser considerar un entorno $(x - h, x + h)$ y repetir el procedimiento anterior.

- Elegimos $h > 0$ y un punto x para clasificar.
- El clasificador identifica en el intervalo $(x - h, x + h)$ los puntos con etiqueta 1 y 0
- Estima a $P(Y = 1 \mid X = x)$ usando un núcleo K

$$\widehat{\mathbb{P}}(Y = 1 \mid X = x) = \frac{\sum_{i=1}^{n} y_i \, K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

- $\widehat{g}_h$
- $h =$?

g^{op} Optimo: Regla de Bayes - Caso binario

Si $X|Y=0 \sim f_0$ y $X|Y=1 \sim f_1$,

resulta

$$g^{op}(x) = \begin{cases} 1 & \text{si } f_1(x)\,\mathbb{P}(Y=1) > f_0(x)\,\mathbb{P}(Y=0) \\ 0 & \text{si c. c.} \end{cases}$$

Estimación g^{op} Optimo: Modelos generativos

$$g^{op}(x) = \begin{cases} 1 & \text{si } f_1(x)\,\mathbb{P}(Y=1) > f_0(x)\,\mathbb{P}(Y=0) \\ 0 & \text{si c. c.} \end{cases}$$

Se estiman

- f_0 y f_1
- $\mathbb{P}(Y=1)$ o $\mathbb{P}(Y=0)$.

Regla plug–in de Bayes

- (X, Y) vector aleatorio
- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- Estimaciones de p_1 y p_0: $\widehat{p}_1 = \frac{n_1}{n}$ y $\widehat{p}_0 = \frac{n_0}{n}$, donde $n_i = \#\{y_i = i\}, i = 0, 1$.

Regla plug–in de Bayes

- (X, Y) vector aleatorio
- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Estimaciones de p_1 y p_0: $\widehat{p}_1 = \frac{n_1}{n}$ y $\widehat{p}_0 = \frac{n_0}{n}$, donde $n_i = \#\{y_i = i\}, i = 0, 1$.
- Estimamos $f_1(x)$ mediante $\widehat{f}_1(x)$.
- Estimamos $f_0(x)$ mediante $\widehat{f}_0(x)$.
- Hacemos un plug–in en la regla $g^{op}(x)$:

$$\widehat{g}(x) = \begin{cases} 1 & \text{si } \widehat{f}_1(x)\,\widehat{p}_1 >= \widehat{f}_0(x)\,\widehat{p}_0, \\ 0 & \text{si } c.c. \end{cases}$$

¿Cómo obtenemos las densidades estimadas $\widehat{f}_1(x)$ y $\widehat{f}_0(x)$?

¿Cómo obtenemos las densidades estimadas $\widehat{f}_1(x)$ y $\widehat{f}_0(x)$?

Opción 1: Enfoque No Paramétrico

- Estimamos $f_1(x)$ no paramétricamente con $\widehat{f}_{1,h_1}(x)$
- Estimamos $f_0(x)$ no paramétricamente con $\widehat{f}_{0,h_0}(x)$
- Hacemos un plug–in en la regla $g^{op}(x)$:

$$\widehat{g}_{h_0,h_1}(x) = \begin{cases} 1 & \text{si } \widehat{f}_{1,h_1}(x)\,\widehat{p}_1 >= \widehat{f}_{0,h_0}(x)\,\widehat{p}_0, \\ 0 & \text{si } c.c. \end{cases}$$

- $\widehat{g}_{h_0,h_1}$, $h_0, h_1 =?$

¿Cómo elegimos k, h o (h_0, h_1)?

En cada caso, tenemos una $\widehat{g}_t$, ¿cómo elegimos el parámetro o vector de parámetros t?

¿Cómo elegimos k, h o (h_0, h_1)?

En cada caso, tenemos una $\widehat{g}_t$, ¿cómo elegimos el parámetro o vector de parámetros t?

Splitting the data

do as well as possible within a given class of rules... This is achieved by splitting the data into a training sequence and a testing sequence. (DGL)

Test Error

Podemos repetir el mismo razonamiento que hicimos para regresión:

- $\widehat{r}_{\mathcal{D}_n}$: $\widehat{r}$ estimador construido a partir de $\mathcal{D}_n$

- Test Error

$$\text{Error}_{\mathcal{D}_n} = \mathbb{E}\left[\{Y_{\text{new}} - \widehat{r}_{\mathcal{D}_n}(\mathbf{X}_{\text{new}})\}^2 \mid \mathcal{D}_n\right]$$

Test Error

Podemos repetir el mismo que hicimos para regresión:

- $\widehat{g}_{\mathcal{D}_n}$: $\widehat{g}$ regla construida a partir de $\mathcal{D}_n$

- Test Error

$$\text{Error}_{\mathcal{D}_n} = \mathbb{E}\left[\{Y_{\text{new}} - \widehat{g}_{\mathcal{D}_n}(\mathbf{X}_{\text{new}})\}^2 \mid \mathcal{D}_n\right]$$

Test Error

Podemos repetir el mismo que hicimos para regresión:

- $\widehat{g}_{\mathcal{D}_n}$: $\widehat{g}$ regla construida a partir de $\mathcal{D}_n$
- Test Error

$$\text{Error}_{\mathcal{D}_n} = \mathbb{E}\left[\{Y_{\text{new}} - \widehat{g}_{\mathcal{D}_n}(\mathbf{X}_{\text{new}})\}^2 \mid \mathcal{D}_n\right]$$

- equivalente a

$$\mathbb{P}\left(\widehat{g}_{\mathcal{D}_n}(\mathbf{X}_{\text{new}}) \neq Y_{\text{new}} \mid \mathcal{D}_n\right)$$

Data-rich situation

- The training set is used to fit the models
- The validation set is used for model selection - con esto elijo tunning parameters.
- The test set is used for assessment of the generalization error of the final chosen model - Acá compiten el mejor candidato de cada posible **método**.

Menos Datos

- Training - Cross Validation.
- The test set is used for assessment of the generalization error of the final chosen model - Acá compiten el mejor candidato de cada posible **método**.

Cross Validation: Leave one out - Fórmulas

t: tunning parameter

$$\mathsf{CV}(t) = \widehat{L(t)} = \frac{1}{n}\sum_{i=1}^{n} L_i(t)$$

- Regresión:

$$L_i(t) = \{Y_i - \widehat{r}_t^{(-i)}(\mathbf{X}_i)\}^2$$

$$t_{\mathsf{opt}} = \mathsf{argmin}\widehat{L(t)}$$

- Clasificación:

$$L_i(t) = \{Y_i - \widehat{g}_t^{(-i)}(\mathbf{X}_i)\}^2 = \mathbb{I}_{\{\widehat{g}_t^{(-i)}(\mathbf{X}_j)\neq Y_j\}}$$

Cross Validation: K folders - Fórmulas

t: tunning parameter

$$\widehat{L(t)} = \frac{1}{K}\sum_{k=1}^{K} L_k(t)$$

- Regresión:

$$L_k(t) = \frac{1}{|\mathcal{T}_k^c|}\sum_{j\in\mathcal{T}_k^c}\{Y_j - \widehat{r}_{t,\mathcal{T}_k}(\mathbf{X}_i)\}^2$$

$$t_{\text{opt}} = \text{argmin}\widehat{L(t)}$$

- Clasificación:

$$L_k(t) = \frac{1}{|\mathcal{T}_k^c|}\sum_{j\in\mathcal{T}_k^c}\mathbb{I}_{\widehat{g}_{t,\mathcal{T}_k}(\mathbf{X}_j)\neq Y_j}$$

Volviendo al método generativo...

- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Estimaciones de p_1 y p_0: $\widehat{p}_1 = \frac{n_1}{n}$ y $\widehat{p}_0 = \frac{n_0}{n}$, donde $n_i = \#\{y_i = i\}, i = 0, 1$.
- Estimamos $f_1(x)$ mediante $\widehat{f}_1(x)$.
- Estimamos $f_0(x)$ mediante $\widehat{f}_0(x)$.
- Hacemos un plug–in en la regla $g^{op}(x)$:

$$\widehat{g}(x) = \begin{cases} 1 & \text{si } \widehat{f}_1(x)\, \widehat{p}_1 >= \widehat{f}_0(x)\, \widehat{p}_0, \\ 0 & \text{si } c.c. \end{cases}$$

¿Cómo obtenemos las densidades estimadas $\widehat{f_1}(x)$ y $\widehat{f_2}(x)$?

¿Cómo obtenemos las densidades estimadas $\widehat{f}_1(x)$ y $\widehat{f}_2(x)$?

Opción 2: Enfoque Paramétrico

Por ejemplo, si fuera razonable asumir que la distribución de la altura X en **cada género** es normal, podríamos estimar los parámetros de cada normal con los datos de altura dentro de cada géneroy hacer un plug–in en la regla de Bayes:

¿Cómo obtenemos las densidades estimadas $\widehat{f}_1(x)$ y $\widehat{f}_2(x)$?

Opción 2: Enfoque Paramétrico

Por ejemplo, si fuera razonable asumir que la distribución de la altura X en **cada género** es normal, podríamos estimar los parámetros de cada normal con los datos de altura dentro de cada géneroy hacer un plug–in en la regla de Bayes:

- Obtenemos $\widehat{\mu}_1$ y $\widehat{\sigma}_1^2$ a partir del género 1: $\Rightarrow \widehat{f}_1(x) = f(x, \widehat{\mu}_1, \widehat{\sigma}_1^2)$.
- Obtenemos $\widehat{\mu}_0$ y $\widehat{\sigma}_0^2$ a partir del género 0: $\Rightarrow \widehat{f}_0(x) = f(x, \widehat{\mu}_0, \widehat{\sigma}_0^2)$.
- Hacemos un plug–in en la regla $g^{op}(x)$:

$$\widehat{g}(x) = \begin{cases} 1 & \text{si } f(x, \widehat{\mu}_1, \widehat{\sigma}_1^2)\, \widehat{p}_1 \geq f(x, \widehat{\mu}_0, \widehat{\sigma}_0^2)\, \widehat{p}_0, \\ 0 & \text{si } c.c. \end{cases}$$

X de mayor dimensión: ingredientes

Supongamos que disponemos de un conjunto de elementos que pueden venir de dos poblaciones distintas.

En cada elemento se ha observado una X variable aleatoria de dimensión p. (por ejemplo registramos altura, ancho, peso, etc.

Se desea clasificar un nuevo elemento, con valores de las variables conocidas.

X de mayor dimensión

Supongamos que para predecir el género del individuo tenemos en cuenta $p = 4$ medidas: altura, peso, perímetro de caderas y de cintura.
$(Y, \mathbf{X})$: donde $\mathbf{X} = (X_1, X_2, \ldots, X_p)^t$, es decir, se registran p covariables.

Igual que antes el clasificador óptimo

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & \text{si } f_1(\mathbf{x})p_1 \geq f_0(\mathbf{x})p_0, \\ 0 & \text{si } c.c. \end{cases}$$

donde $f_1(\mathbf{x})$ y $f_0(\mathbf{x})$ son densidades en dimensión p.

X de mayor dimensión

Opción 1: Enfoque No Paramétrico

Si no sabemos nada sobre las dos densidades, una alternativa es usar el estimador no parámetrico de la densidad al igual que antes
Ahora tenemos una dendidad sobre $\mathbb{R}^p$: $d(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^p$.
Dada la

$$\widehat{d}(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right)$$

ahora $\mathbf{x} - \mathbf{X}_i$ es de dimensión p.

Bayes Naive

Una forma de eludir el problema de la maldición de la dimensión es construir un estimador bajo la hipótesis de que las p covariables son independientes.

Bayes Naive

Una forma de eludir el problema de la maldición de la dimensión es construir un estimador bajo la hipótesis de que las p covariables son independientes. Luego, pensamos que

$$d(\mathbf{x}) = d_1(x_1)\, d_2(x_2) \ldots d_p(x_p)$$

y estimamos:

$$\widehat{d}(\mathbf{x}) = \widehat{d}_1(x_1)\, \widehat{d}_2(x_2) \ldots \widehat{d}_p(x_p)$$

Bayes Naive

Una forma de eludir el problema de la maldición de la dimensión es construir un estimador bajo la hipótesis de que las p covariables son independientes. Luego, pensamos que

$$d(\mathbf{x}) = d_1(x_1)\, d_2(x_2) \ldots d_p(x_p)$$

y estimamos:

$$\widehat{d}(\mathbf{x}) = \widehat{d}_1(x_1)\, \widehat{d}_2(x_2) \ldots \widehat{d}_p(x_p)$$

Es decir, en lugar de estimar la densidad multivariada de $\mathbf{X}$ =(altura, peso, perímetro de caderas, perímetro de cintura). Estimamos de forma separada la densidad de X_1 = altura, X_2 = peso, X_3 = perímetro de caderas y X_4 = perímetro de cintura.

Bayes Naive

Una forma de eludir el problema de la maldición de la dimensión es construir un estimador bajo la hipótesis de que las p covariables son independientes. Luego, pensamos que

$$d(\mathbf{x}) = d_1(x_1)\, d_2(x_2) \ldots d_p(x_p)$$

y estimamos:

$$\widehat{d}(\mathbf{x}) = \widehat{d}_1(x_1)\, \widehat{d}_2(x_2) \ldots \widehat{d}_p(x_p)$$

Es decir, en lugar de estimar la densidad multivariada de $\mathbf{X}$ =(altura, peso, perímetro de caderas, perímetro de cintura). Estimamos de forma separada la densidad de X_1 = altura, X_2 = peso, X_3 = perímetro de caderas y X_4 = perímetro de cintura. Así, obtenemos $\widehat{f}_1(\mathbf{x})$ y $\widehat{f}_0(\mathbf{x})$ y con esto armamos el clasificador.

$$\widehat{g}(\mathbf{x}) = \begin{cases} 1 & \text{si } \widehat{f}_1(\mathbf{x})\widehat{p}_1 \geq \widehat{f}_0(\mathbf{x})\widehat{p}_0, \\ 0 & \text{si } c.c. \end{cases}$$

Otros métodos de clasificación

Opción 2: Enfoque Paramétrico

Revisitando Vectores Aleatorios

Sea $\mathbf{X}$ es un vector aleatorio p-dimensional tal que

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix},$$

donde las componentes $X_1, \ldots, X_p$ son variables aleatorias.
Su **vector de medias** es

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix},$$

donde μ_i es la esperanza de la componente X_i.

Covarianza

Dadas dos variables aleatorias U y V con esperanza μ_U y μ_V llamamos covarianza entre U y V a

$$\mathbb{C}\text{ov}(U, V) \quad = \quad \mathbb{E}\{(U - \mu_U)(V - \mu_V)\}$$

Covarianza

Dadas dos variables aleatorias U y V con esperanza μ_U y μ_V llamamos covarianza entre U y V a

$$\begin{aligned} \mathbb{C}\text{ov}(U,V) &= \mathbb{E}\{(U-\mu_U)(V-\mu_V)\} \\ &= \mathbb{E}\{UV\} - \mu_U\mu_V \\ &= \sigma_{UV} \end{aligned}$$

Covarianza

Dadas dos variables aleatorias U y V con esperanza μ_U y μ_V llamamos covarianza entre U y V a

$$\begin{aligned} \mathbb{C}\text{ov}(U,V) &= \mathbb{E}\{(U-\mu_U)(V-\mu_V)\} \\ &= \mathbb{E}\{UV\} - \mu_U\mu_V \\ &= \sigma_{UV} \end{aligned}$$

Notemos que

1. $\mathbb{C}\text{ov}(U,U) = \mathbb{V}ar(U)$.
2. Si U y V son independientes $\Rightarrow \mathbb{C}\text{ov}(U,V) = 0$.
3. No vale la recíproca.
4. $\mathbb{C}\text{ov}(aU,bV) = ab\ \mathbb{C}\text{ov}(U,V)$

Correlación

Dadas dos variables aleatorias U y V con desvío standard σ_U y σ_V, la correlación entre U y V es

$$\rho \quad = \quad \frac{\mathbb{C}\text{ov}(U, V)}{\sigma_U \ \sigma_V}$$

"mide la asociación lineal" entre U y V.

Correlación

Dadas dos variables aleatorias U y V con desvío standard σ_U y σ_V, la correlación entre U y V es

$$\rho \quad = \quad \frac{\mathbb{C}\text{ov}(U, V)}{\sigma_U \; \sigma_V}$$

"mide la asociación lineal" entre U y V.

- $-1 \leq \rho \leq 1$.
- Si $|\rho| = 1$, entonces $V = \alpha + \beta U$ para algún α y β.
- $\mathbb{C}\text{ov}(U, V) = \rho \; \sigma_U \; \sigma_V$.

Matriz de Covarianza: $\mathbf{X} = (X_1, X_2, \ldots, X_p)$

Matriz de Covarianza de $\mathbf{X}$: $\mathbf{\Sigma}_{ij} = \mathbb{C}\mathsf{ov}(X_i, X_j)$

Matriz de Covarianza: $\mathbf{X} = (X_1, X_2, \ldots, X_p)$

Matriz de Covarianza de $\mathbf{X}$: $\mathbf{\Sigma}_{ij} = \mathbb{C}\mathsf{ov}(X_i, X_j)$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbb{V}\mathrm{ar}(X_1) & \mathbb{C}\mathsf{ov}(X_1, X_2) & \ldots & \mathbb{C}\mathsf{ov}(X_1, X_p) \\ \mathbb{C}\mathsf{ov}(X_2, X_1) & \mathbb{V}\mathrm{ar}(X_2) & \ldots & \mathbb{C}\mathsf{ov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathsf{ov}(X_p, X_1) & \mathbb{C}\mathsf{ov}(X_p, X_2) & \ldots & \mathbb{V}\mathrm{ar}(X_p) \end{pmatrix}$$

Matriz de Covarianza: $\mathbf{X} = (X_1, X_2, \ldots, X_p)$

Matriz de Covarianza de $\mathbf{X}$: $\mathbf{\Sigma}_{ij} = \mathbb{C}\mathsf{ov}(X_i, X_j) = \sigma_{ij}$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Matriz de Covarianza: $\mathbf{X} = (X_1, X_2)$

Matriz de Covarianza en dimensión $p = 2$:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

Matriz de Covarianza: $\mathbf{X} = (X_1, X_2)$

Matriz de Covarianza en dimensión $p = 2$:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\, \sigma_1\sigma_2 \\ \rho\, \sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

donde $\sigma_{ij} = \mathbb{C}\text{ov}(X_i, X_j)$ y ρ es la correlación.

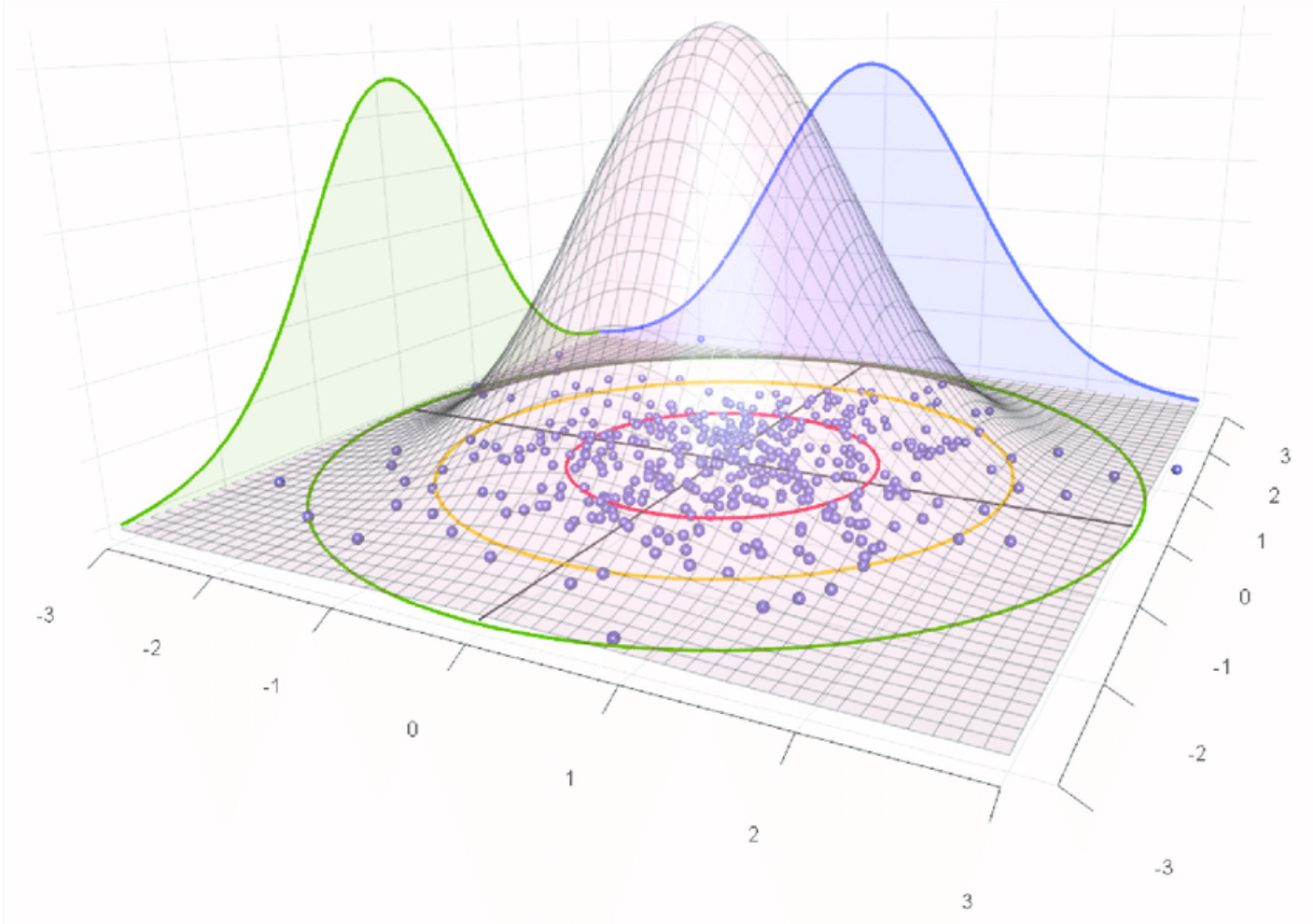
Normal multivariada

Diremos que X es un vector aleatorio con distribución normal multivariada en R^p, que denotaremos $N_p(\mu, \Sigma)$ si su densidad es

$$f_X(x) = \frac{1}{2\pi \left[\det(\Sigma)\right]^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right\}.$$

Normal multivariada

Por ejemplo la función de densidad de una variables $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ en dimensión 2

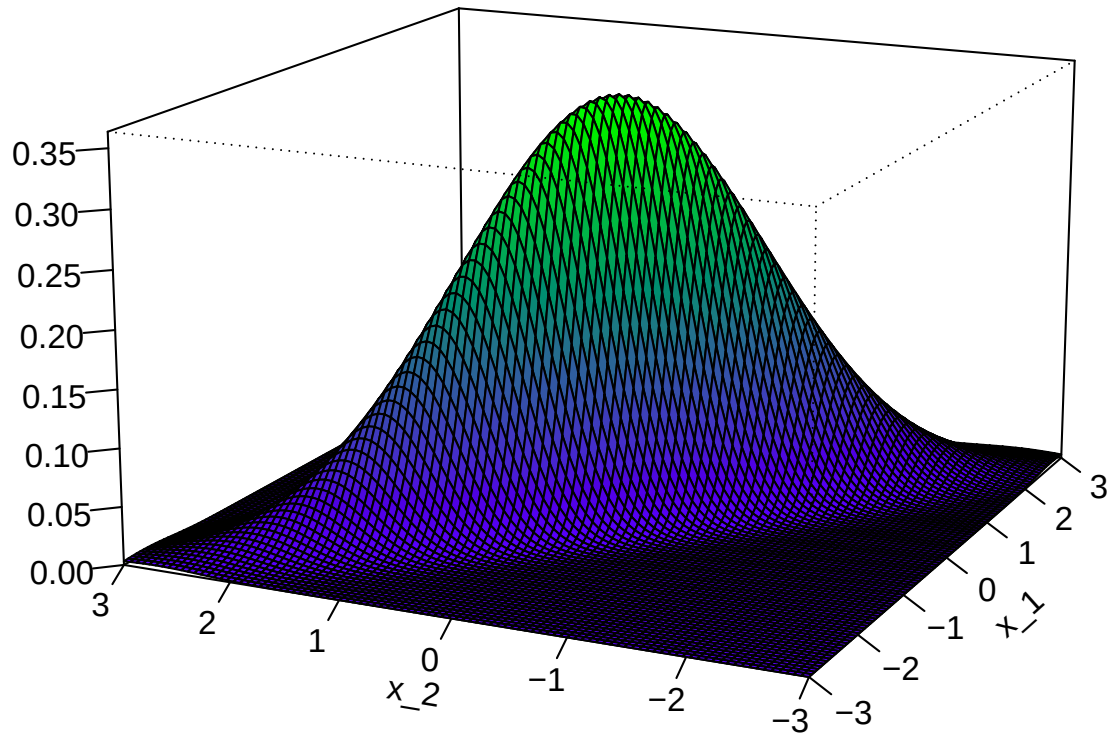


Caso $p = 2$

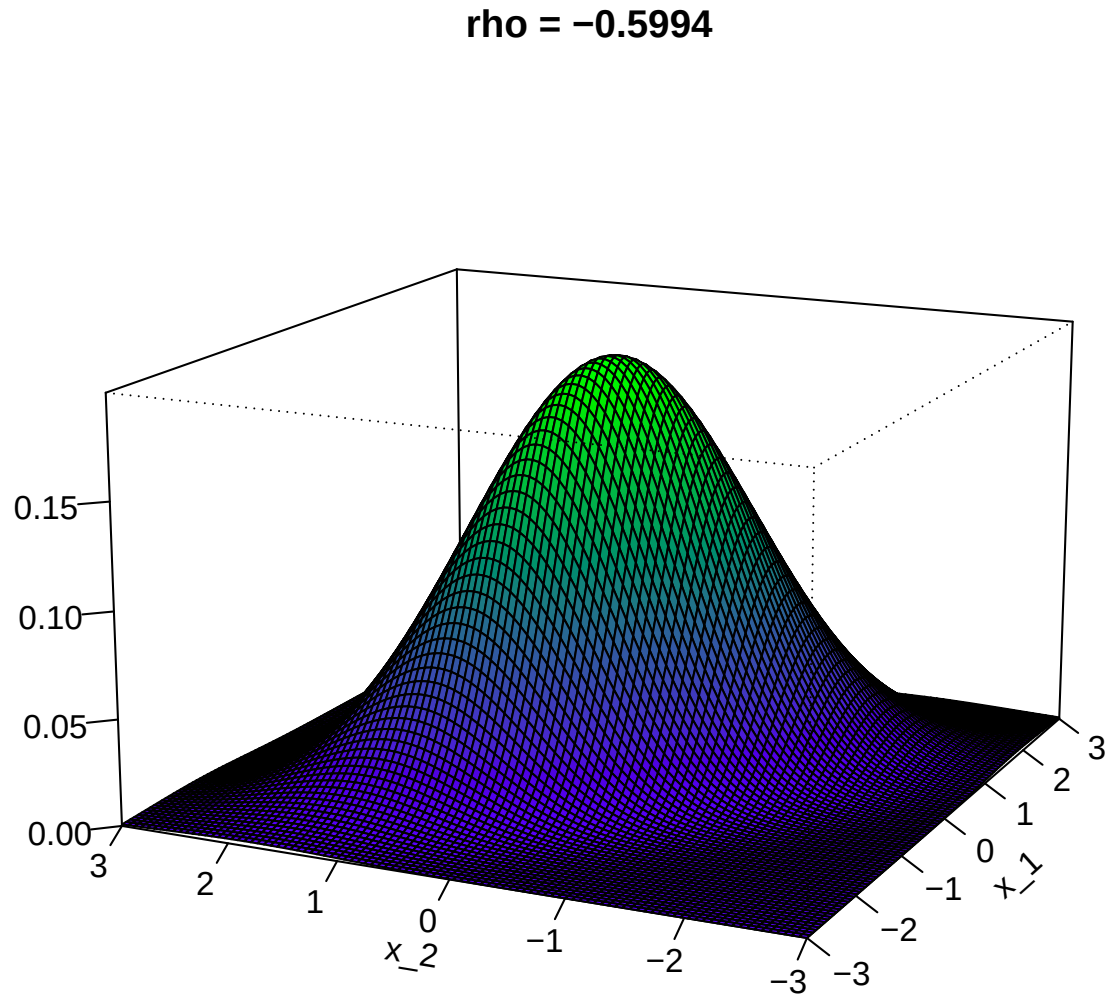
- Sea $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^2$ y $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ definida positiva ($|\rho| \neq 1$)

$$f(\mathbf{x}) = \frac{1}{2\pi} \frac{1}{\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) \right] \right\}$$

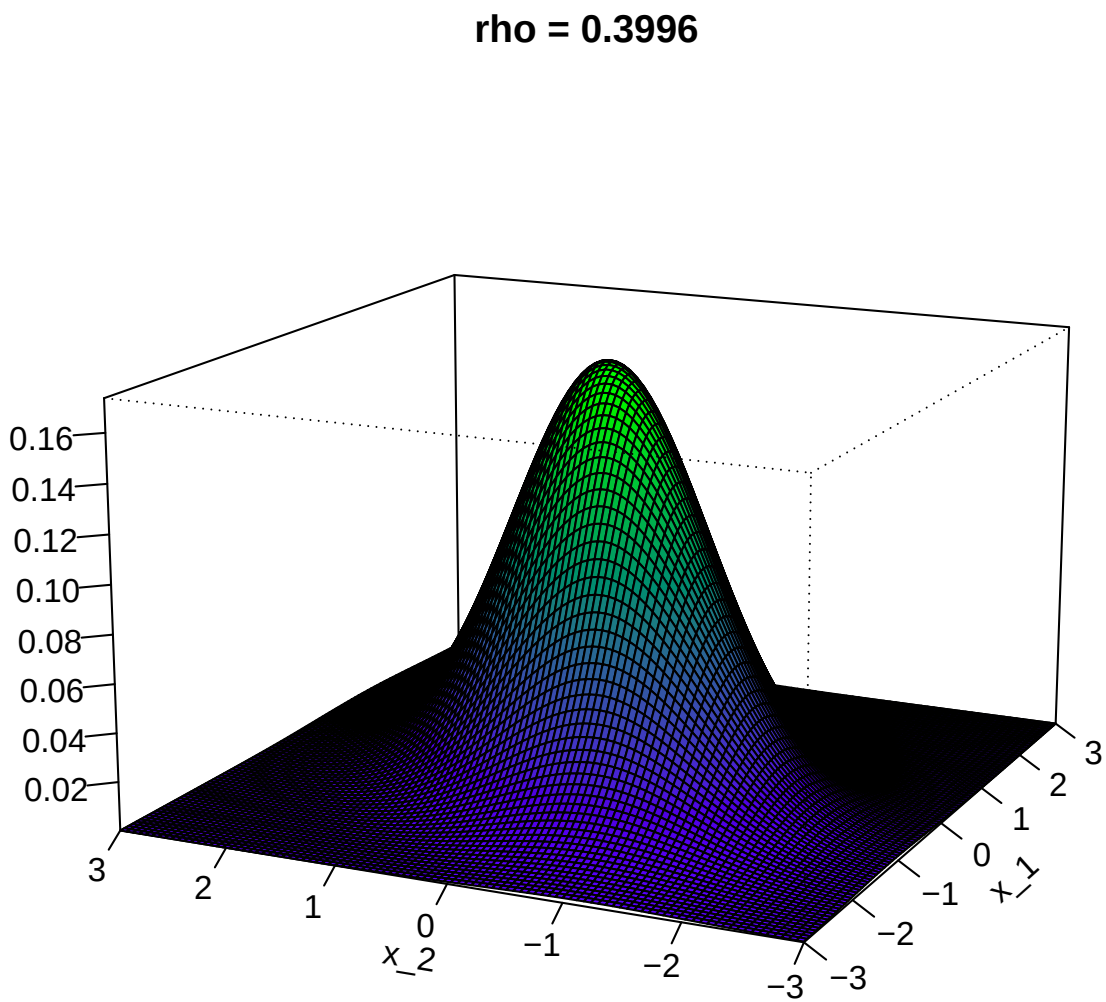
Caso p=2



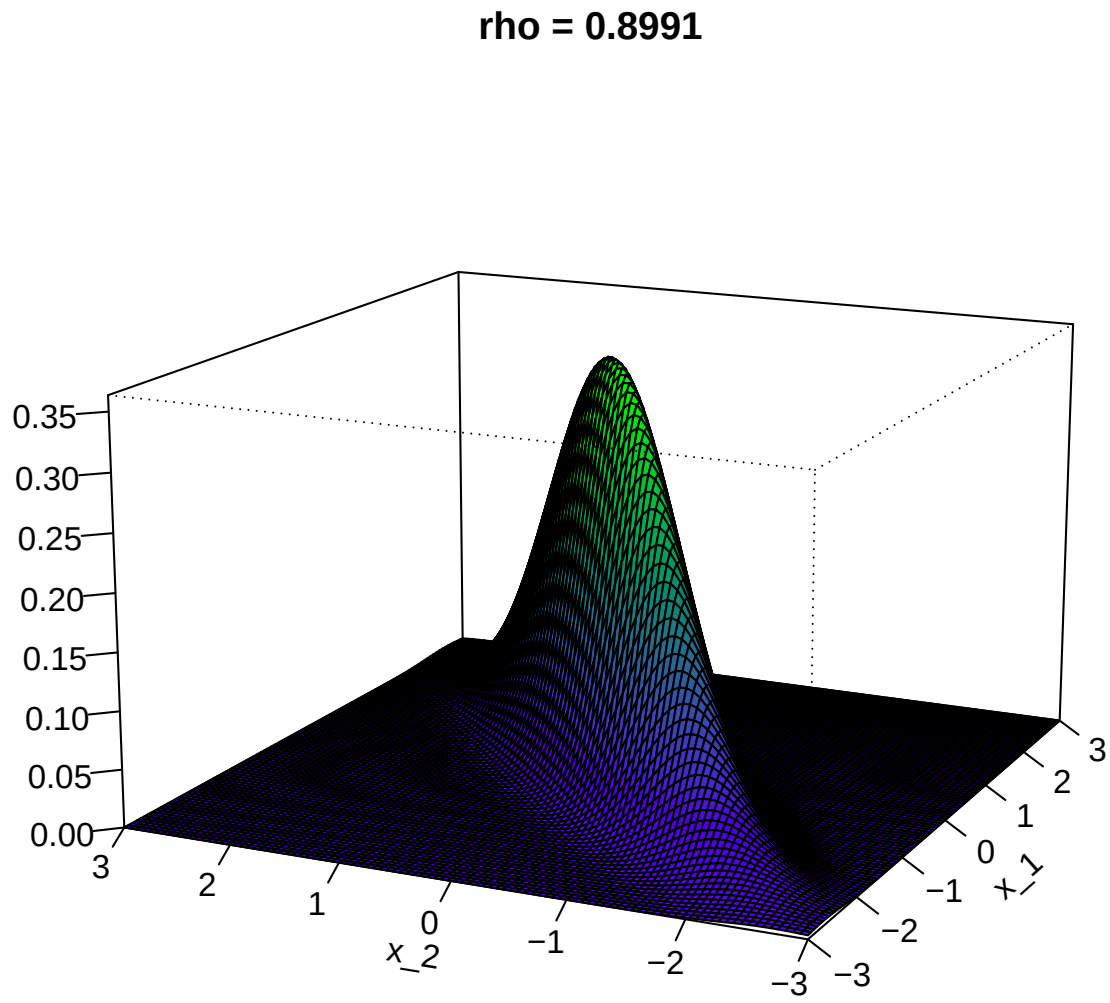
Caso p=2



Caso p=2



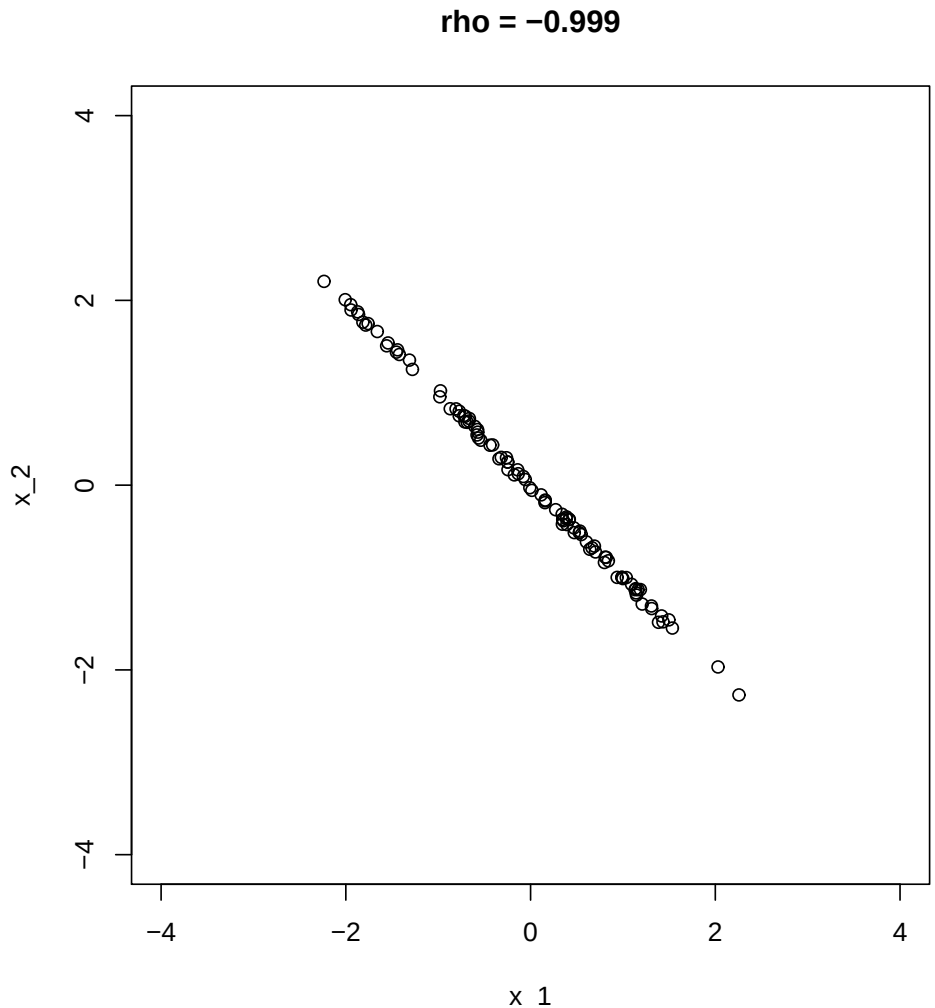
Caso p=2



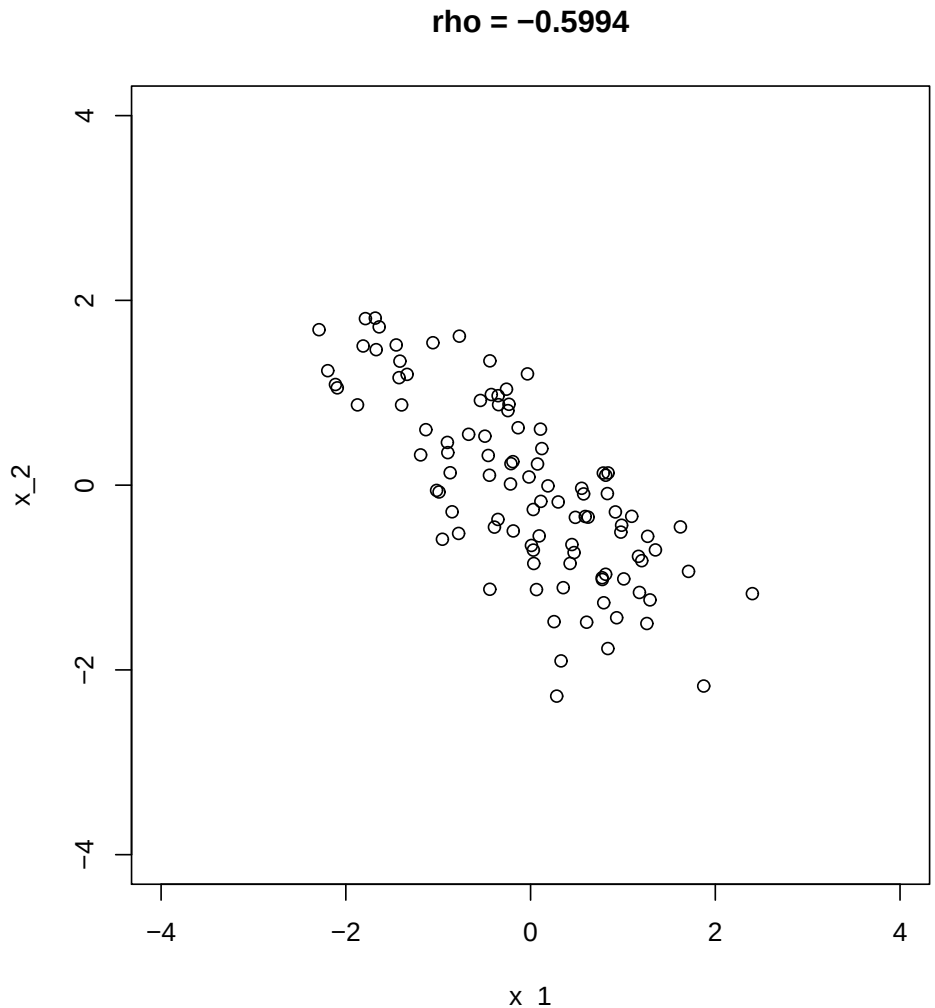
Caso p=2

Ahora vamos a muestrear de estas distribuciones.

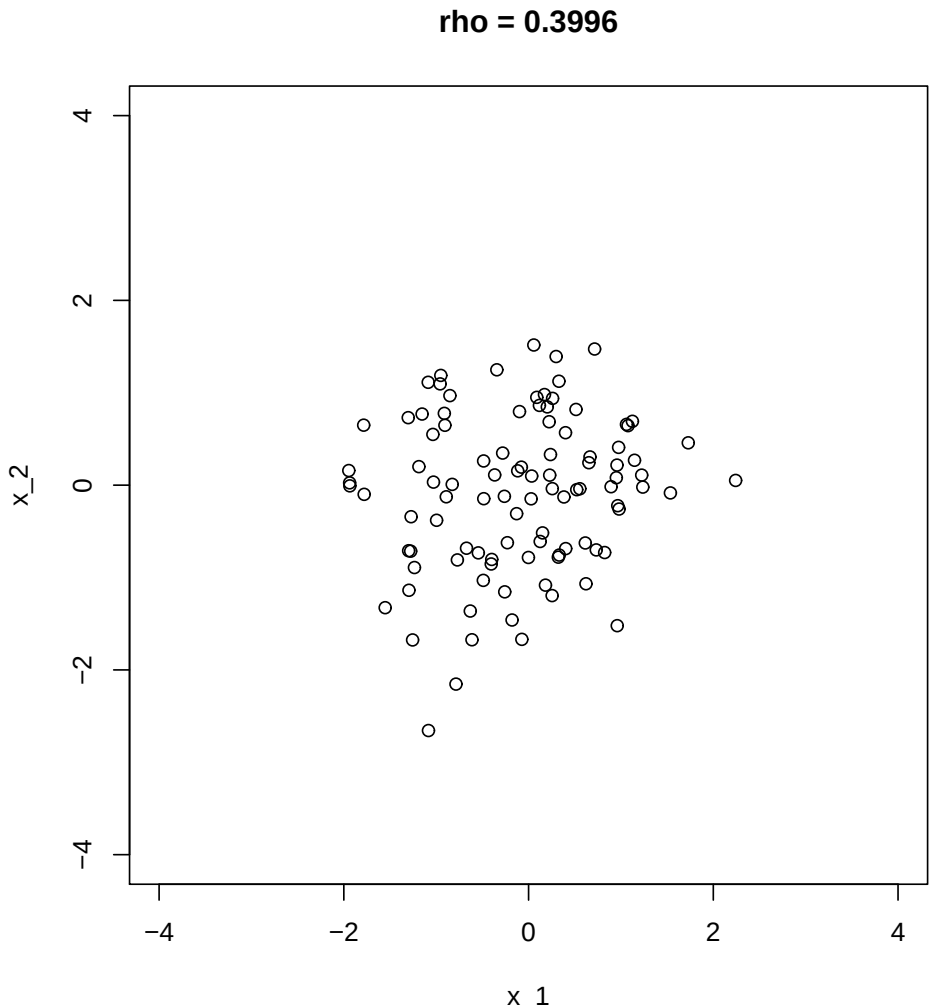
Caso p=2



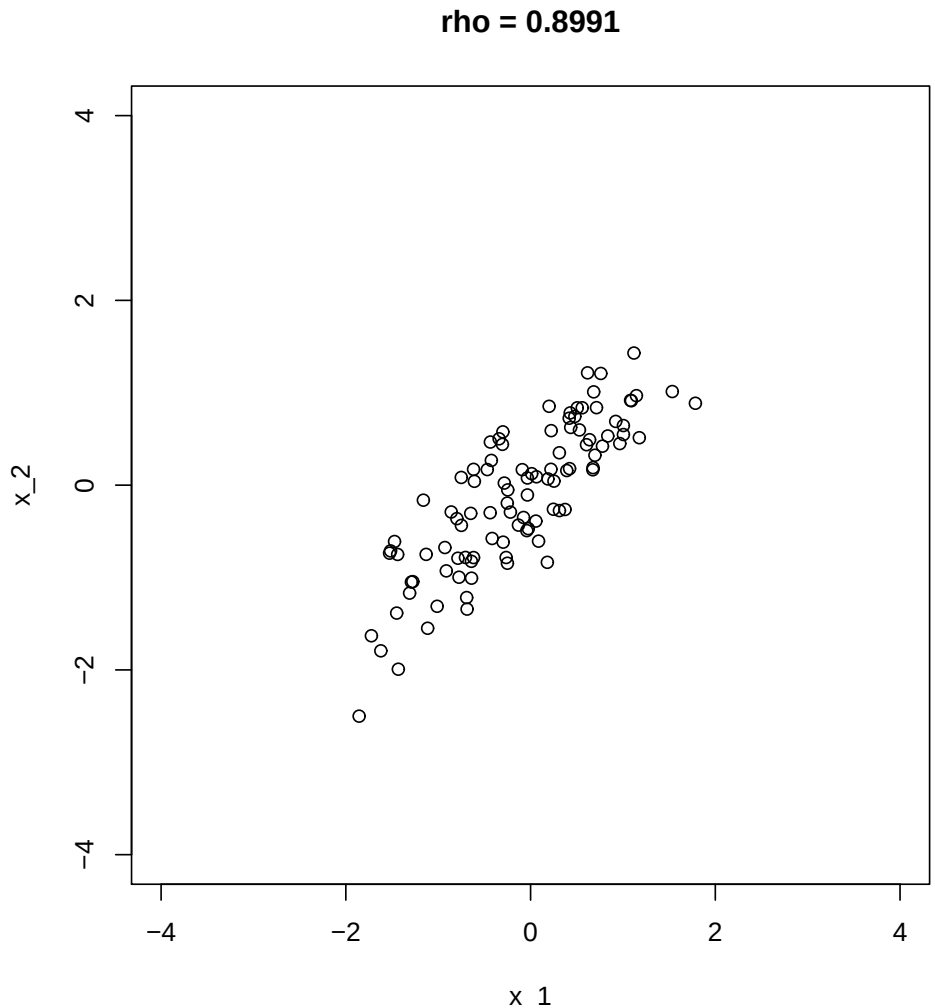
Caso p=2



Caso p=2



Caso p=2



Distancia de Mahalanobis

- La distancia de Mahalanobis de un punto $\mathbf{x}$ a la media $\boldsymbol{\mu}$ es D siendo

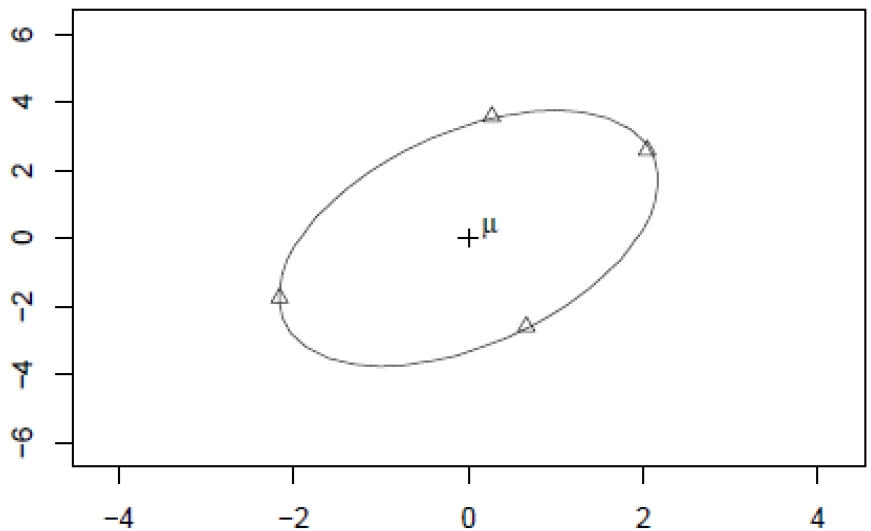
$$D^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Distancia de Mahalanobis

- La distancia de Mahalanobis de un punto $\mathbf{x}$ a la media $\boldsymbol{\mu}$ es D siendo

$$D^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- De esta forma, dos puntos tienen la misma distancia de Mahalanobis si están en el mismo elipsoide centrado en $\boldsymbol{\mu}$



¿Cómo estimamos?

Supongamos que tenemos una muestra de vectores aleatorios $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ independientes y $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Es decir, ahora observamos vectores aleatorios que constituyen una matriz de observaciones de $n \times p$:

$$\begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{22} & \ldots & X_{np} \end{pmatrix}$$

Los estimadores de máxima verosimilitud de la media y la matriz de covarianzas son

- $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$
- $\widehat{\boldsymbol{\Sigma}} = \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^t$

Clasificación: Discriminación Lineal

Volvamos al problema de las alturas y la predicción del género:

- Dos clases: género F (1) y género M (0)
  - p_1 es la proporción de individuos de género 1
  - $p_0 = 1 - p_1$ de individuos de género 0.

- Estamos interesados en determinar el género de un individuo a partir de su (altura,peso,...)... p características $\mathbf{X}$
  - en el género 1 $\mathbf{X}$ es un v.a con densidad $f_1(\mathbf{x})$ normal multivariada $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
  - en el género 0 $\mathbf{X}$ es un v.a con densidad $f_0(\mathbf{x})$ normal multivariada $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$

Clasificación: Discriminación Lineal

Volvamos al problema de las alturas y la predicción del género:

- Dos clases: género F (1) y género M (0)
  - p_1 es la proporción de individuos de género 1
  - $p_0 = 1 - p_1$ de individuos de género 0.

- Estamos interesados en determinar el género de un individuo a partir de su (altura,peso,...)... p características $\mathbf{X}$
  - en el género 1 $\mathbf{X}$ es un v.a con densidad $f_1(\mathbf{x})$ normal multivariada $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
  - en el género 0 $\mathbf{X}$ es un v.a con densidad $f_0(\mathbf{x})$ normal multivariada $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$

Ambas variedades normales multivariadas con distinta media, pero igual matriz de covarianza $\boldsymbol{\Sigma}$.

Clasificación: Discriminación Lineal

Igual que antes el clasificador óptimo

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & \text{si } f_1(\mathbf{x})p_1 \geq f_0(\mathbf{x})p_0, \\ 0 & \text{si } c.c. \end{cases}$$

Clasificación: Discriminación Lineal

Igual que antes el clasificador óptimo

$$g^{op}(\mathbf{x}) = \begin{cases} 1 & \text{si } f_1(\mathbf{x})p_1 \geq f_0(\mathbf{x})p_0, \\ 0 & \text{si } c.c. \end{cases}$$

el clasificador quedaría que clasifica en la variedad 1 si

$$\frac{\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)^t \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)\right\}}{2\pi\left[\det\left(\boldsymbol{\Sigma}\right)\right]^{\frac{1}{2}}} p_1 \geq \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_0\right)^t \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_0\right)\right\}}{2\pi\left[\det\left(\boldsymbol{\Sigma}\right)\right]^{\frac{1}{2}}} p_0$$

Clasificación: Discriminación Lineal

Si consideramos las distancias de Mahalanobis

$$D_1 = (x - \mu_1)^t \, \Sigma^{-1} \, (x - \mu_1) \qquad D_0 = (x - \mu_0)^t \, \Sigma^{-1} \, (x - \mu_0)$$

Clasificación: Discriminación Lineal

Si consideramos las distancias de Mahalanobis

$$D_1 = (x - \mu_1)^t \, \Sigma^{-1} \, (x - \mu_1) \qquad D_0 = (x - \mu_0)^t \, \Sigma^{-1} \, (x - \mu_0)$$

haciendo algunos manejos algebraicos la desigualdad anterior se transforma y el clasificador quedaría que clasifica en la clase 1 si

$$D_1 \leq D_0 - \log(p_0/p_1)$$

Clasificación: Discriminación Lineal

Si cada variedad es una variable $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, entonces se clasifica en la variedad 1 si

$$\mathbf{w}^t \mathbf{x} \leq \mathbf{w}^t \, \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} - \log(p_0/p_1)$$

con $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

Clasificación: Discriminación Lineal

Si cada variedad es una variable $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, entonces se clasifica en la variedad 1 si

$$\mathbf{w}^t \mathbf{x} \leq \mathbf{w}^t \, \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} - \log(p_0/p_1)$$

con $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

El clasificador corresponde a la ecuación de un hiperplano, es decir hay un hiperplano que separa a quien clasifica en la variedad 1 y a quien en la 0 (de ahí el nombre de clasificador lineal.)

¿Si no conocemos $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ y $\boldsymbol{\Sigma}$?

Clasificación: Discriminación Lineal

Si cada variedad es una variable $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, entonces se clasifica en la variedad 1 si

$$\mathbf{w}^t \mathbf{x} \leq \mathbf{w}^t \, \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} - \log(p_0/p_1)$$

con $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

El clasificador corresponde a la ecuación de un hiperplano, es decir hay un hiperplano que separa a quien clasifica en la variedad 1 y a quien en la 0 (de ahí el nombre de clasificador lineal.)

¿Si no conocemos $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ y $\boldsymbol{\Sigma}$?
los estimamos!
$\widehat{\boldsymbol{\mu}}_1$, $\widehat{\boldsymbol{\mu}}_0$ y $\widehat{\boldsymbol{\Sigma}}_{pooled} = \dfrac{(n_1 - 1)\mathbf{S}_1 + (n_0 - 1)\mathbf{S}_0}{n_1 + n_0 - 2}$

Clasificación: Discriminación Lineal

Las condiciones que se deben cumplir para que un Análisis Discriminante Lineal (LDA) sea válido son:

- Las observaciones siguen una distribución normal multivariada en todas las clases.

- La matriz de covarianza es igual en todas las clases. Si esto no se cumple se recurre a Análisis Discriminante Cuadrático (QDA).

Clasificación: Discriminación Cuadrática

Si cada variedad es una variable $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ y $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, es decir no asumimos igual varianza.

Clasificación: Discriminación Cuadrática

Si cada variedad es una variable $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ y $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, es decir no asumimos igual varianza. Se clasifica en la variedad 1 si

$$(\mathbf{x} - \boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - \log(p_1) \leq (\mathbf{x} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) - \log(p_0)$$

Clasificación: Discriminación Cuadrática

Si cada variedad es una variable $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ y $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, es decir no asumimos igual varianza. Se clasifica en la variedad 1 si

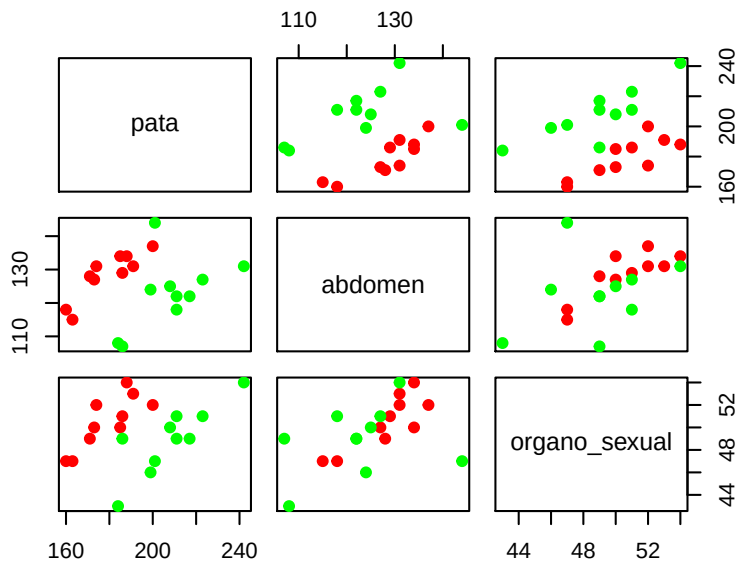
$$(\mathbf{x} - \boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - \log(p_1) \leq (\mathbf{x} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) - \log(p_0)$$

Si graficamos esta región ya no queda una separación lineal entre las clases sino cuadrática.

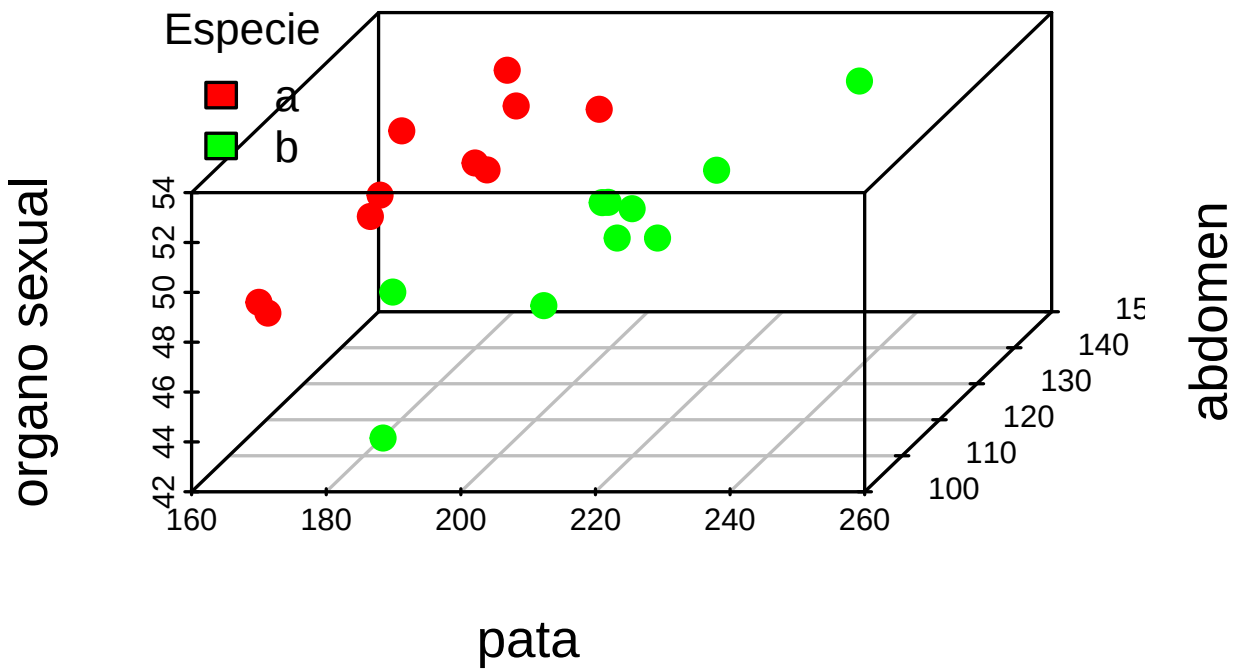
¿Cómo estimamos a $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_1$ y $\boldsymbol{\Sigma}_0$?

Discrinación lineal y cuadrática

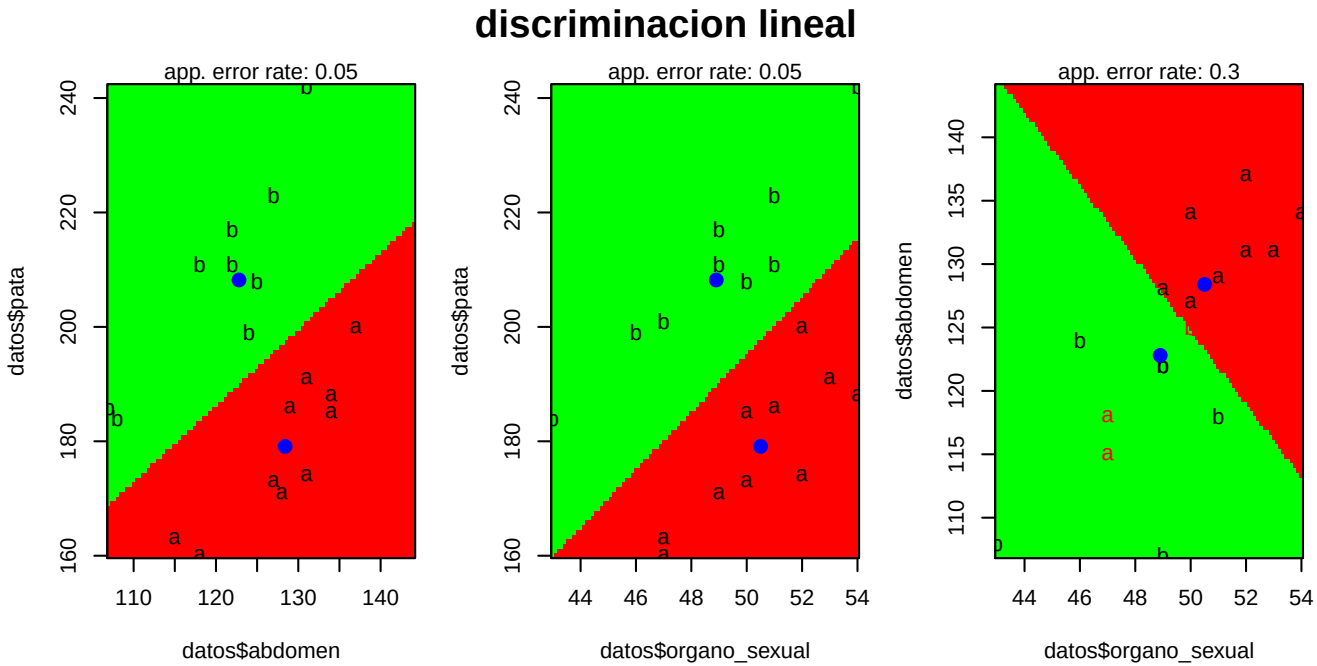
Se quiere generar un modelo estadístico que permita identificar a que especie (a o b) pertenece un determinado insecto.
Se mide $X =$(longitud de las patas, diámetro del abdomen, diámetro del órgano sexual) en 10 individuos de cada una de las dos especies.



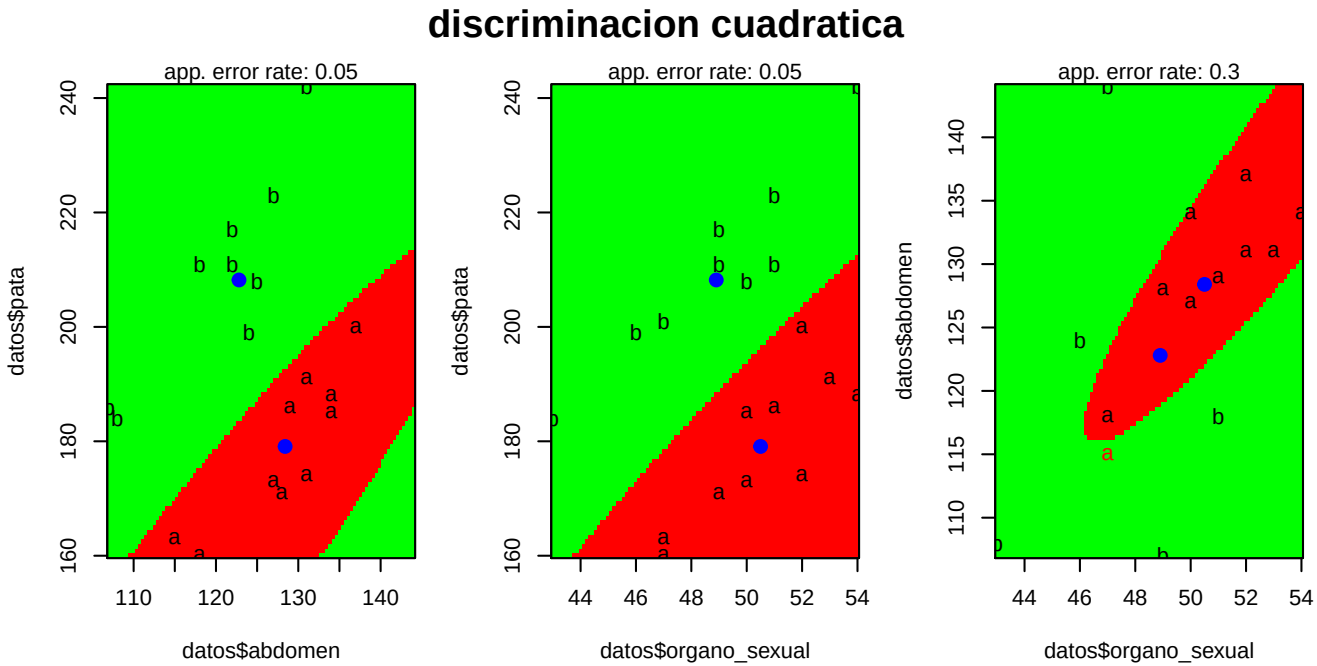
Discriminación lineal y cuadrática



Discriminación lineal y cuadrática



Discriminación lineal y cuadrática



Volvamos al Enfoque Discriminativo....

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 $\mathbf{X}$?

Volvamos al Enfoque Discriminativo....

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...

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**

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

Regresión Logística

Pensemos que tenemos una sola variable X
Notemos que

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

Sin embargo, los **odds**

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

Tomando logaritmo:

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

Podríamos modelar:

$$\log\left(\frac{p(x)}{1-p(x)}\right) \quad = \quad \beta_1 + \beta_2\, x$$

Regresión Logística

Pensemos que tenemos una sola variable X
Notemos que

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

Sin embargo, los **odds**

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

Tomando logaritmo:

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

Podríamos modelar:

$$logit(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) \quad = \quad \beta_1 + \beta_2\, x$$

Regresión Logística

Haciendo el camino inverso:

$$\frac{p(x)}{1-p(x)} = e^{\beta_1+\beta_2\ x}$$

$$p(x) = p(x, \boldsymbol{\beta}) = \frac{e^{\beta_1+\beta_2\ x}}{1+e^{\beta_1+\beta_2\ x}}$$

Regresión Logística

Haciendo el camino inverso:

$$\frac{p(x)}{1-p(x)} = e^{\beta_1+\beta_2\ x}$$

$$p(x) = p(x, \boldsymbol{\beta}) = \frac{e^{\beta_1+\beta_2\ x}}{1+e^{\beta_1+\beta_2\ x}} = \frac{1}{1+e^{-\beta_1-\beta_2\ x}}$$

Regresión Logística

Con una sola variable

$$p(x) = p(x, \boldsymbol{\beta}) \quad = \quad \frac{e^{\beta_1 + \beta_2\ x}}{1 + e^{\beta_1 + \beta_2\ x}}$$

Regresión Logística

Con una sola variable

$$p(x) = p(x, \boldsymbol{\beta}) \quad = \quad \frac{e^{\beta_1 + \beta_2\ x}}{1 + e^{\beta_1 + \beta_2\ x}}$$

En general:

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

Regresión Logística

Con una sola variable

$$p(x) = p(x, \boldsymbol{\beta}) \quad = \quad \frac{e^{\beta_1 + \beta_2\ x}}{1 + e^{\beta_1 + \beta_2\ x}}$$

En general:

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

Estamos proponiendo un modelo para la distribución condicional:

$$Y|_{\mathbf{X}=\mathbf{x}} \sim Bi(1, p(\mathbf{x}, \boldsymbol{\beta}))$$

Regresión Logística: Estimación

$(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ independientes donde $Y_i|_{\mathbf{X}_i=\mathbf{x}} \sim Bi(1, p(\mathbf{x}, \boldsymbol{\beta}))$

El estimador clásico en este contexto se obtiene por el método de máxima verosimilitud, es decir hallando β_1, $\beta_2, \ldots,$ β_p que maximizan $L(\boldsymbol{\beta}) = L(\boldsymbol{\beta}; y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n)$ donde

$$L(\boldsymbol{\beta}; y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} p(\mathbf{x}_i, \boldsymbol{\beta})^{y_i} (1 - p(\mathbf{x}_i, \boldsymbol{\beta}))^{1-y_i}$$

Regresión Logística: Máxima Verosimilitud

$\widehat{\boldsymbol{\beta}}$ resulta de maximizar

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

Regresión Logística: Máxima Verosimilitud

$\widehat{\boldsymbol{\beta}}$ resulta de maximizar

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

o minimizar

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

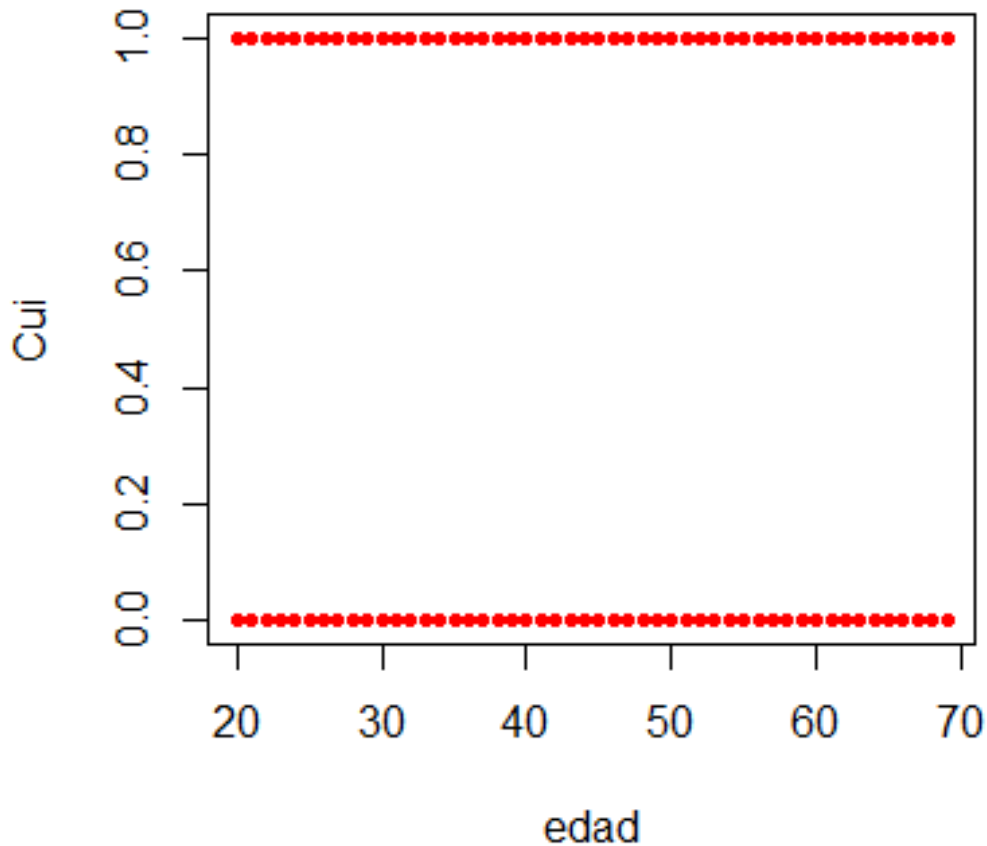
relacionado con la deviance y con la cross–entropy.
Este estimador no tiene una expresión analítica. Se halla derivando e igualando a 0 la log–verosimilitud y resolviendo numéricamente la ecuación por el método de Newton-Raphson o Fisher-scoring.

Vayamos a un ejemplo: datos de COVID

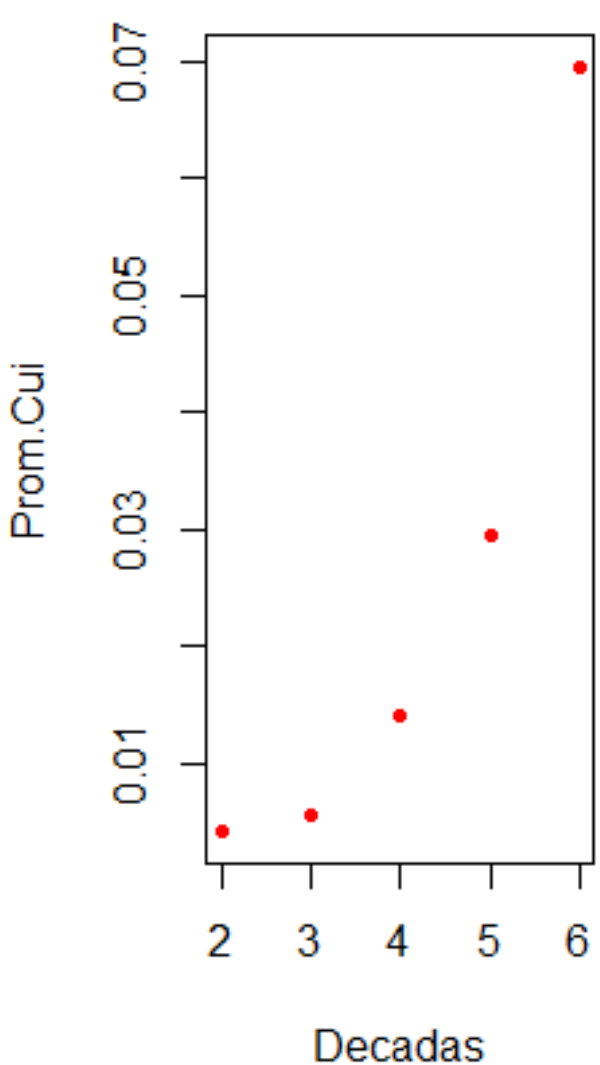
Usamos los datos oficiales suministrados por el Ministerio de Salud hasta el 09–07–2020.
Consideramos los casos confirmados de adultos entre 20 y 70 años y las variables Variables:

- **cui**: cuidados intensivos (1=si o 0=no)
- **edad**
- **género**: F (1) o M (0)

Gráfico: CUI vs. Edad



Gráficos


Prom.Cui
0.01
0.03
0.05
0.07
2
3
4
5
6
Decadas

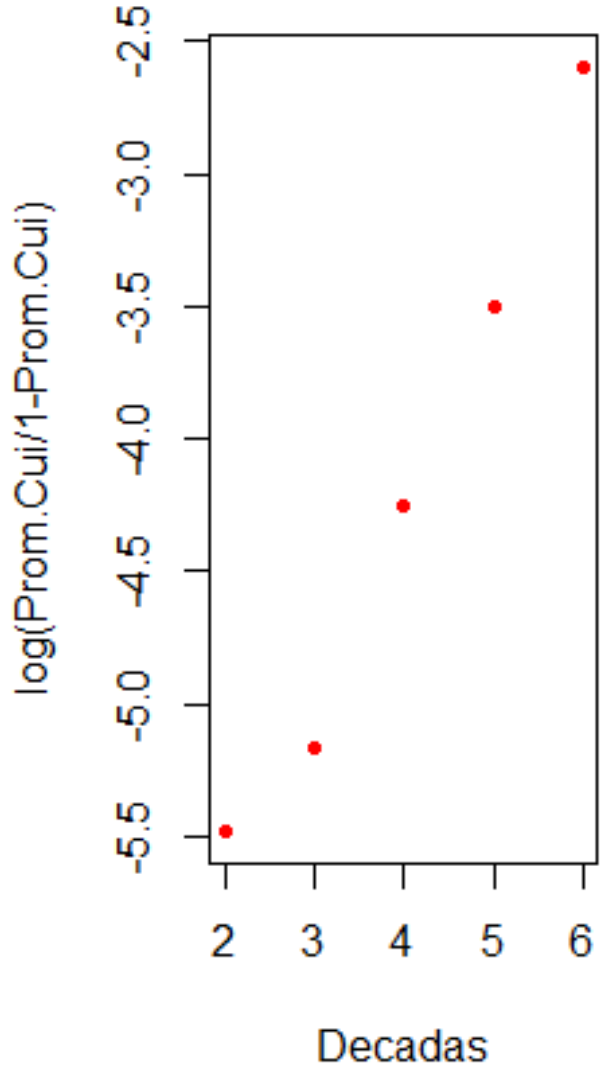
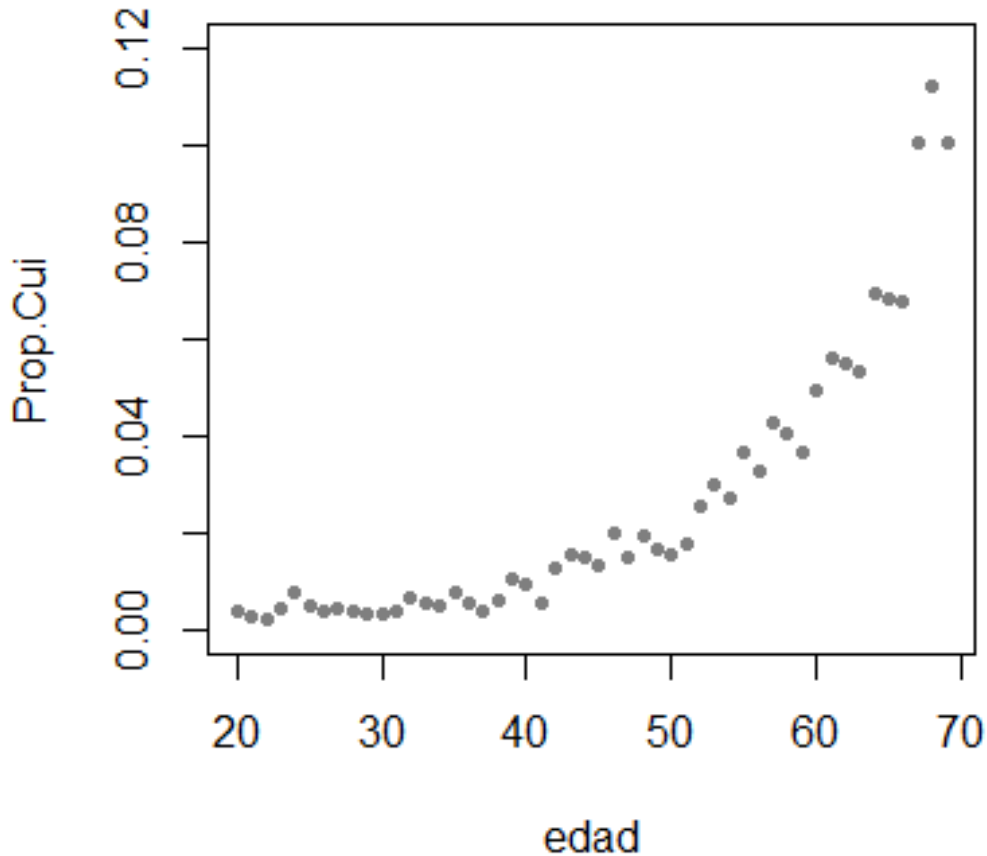

log(Prom.Cui/1-Prom.Cui)
-5.5
-5.0
-4.5
-4.0
-3.5
-3.0
-2.5
2
3
4
5
6
Decadas

Gráfico: proporción de CUI vs. Edad



Algunas medidas de resumen

```
> table(sexo)
sexo
F     M
34251 36731

> table(cuidado_intensivo)
cuidado_intensivo
NO    SI
69853  1129

> summary(edad.dec)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
20.00   29.00   38.00   39.56   49.00   69.00
```

Ajustamos usando Edad

Llamemos

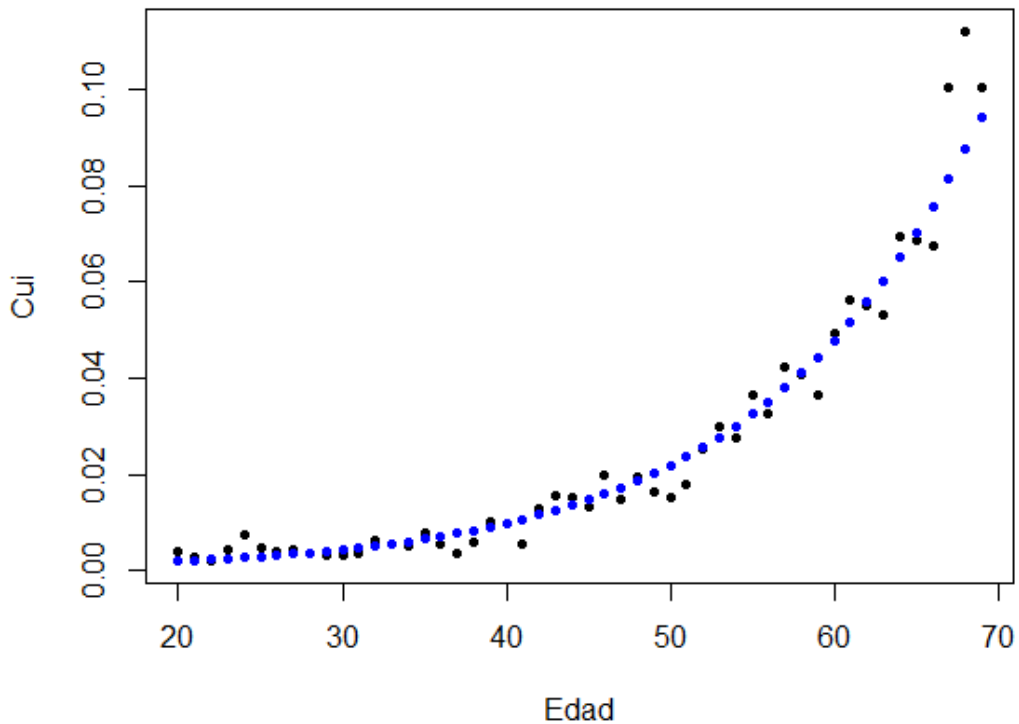
$$p(\text{Edad})) = \mathbb{P}(\text{CUI} = Si \mid \text{Edad})$$

Ajustamos el modelo

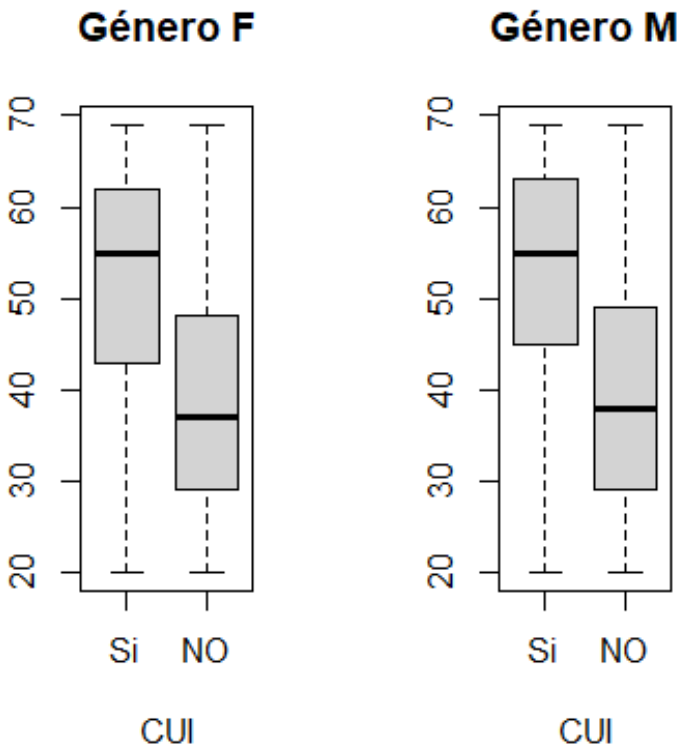
$$logit(p(\text{Edad}_i)) = \beta_1 + \beta_2\ \text{Edad}_i$$

```
proba.hat=glm(cui~edad.dec,family=binomial)$fitted

##########  promedio x genero

promedios<- datos%>%  group_by(edad.dec) %>% summarize(mean(cui))

plot(promedios[[1]],promedios[[2]],xlab = "Edad",ylab = "Cui",pch=20)
points(edad.dec[genero==1],proba.hat[genero==1],col="red",pch=20)
```

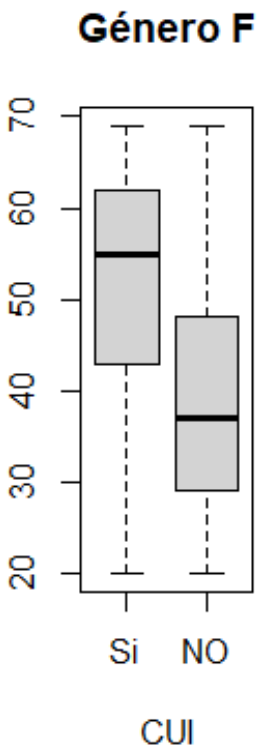
Gráfico: en azul predichos

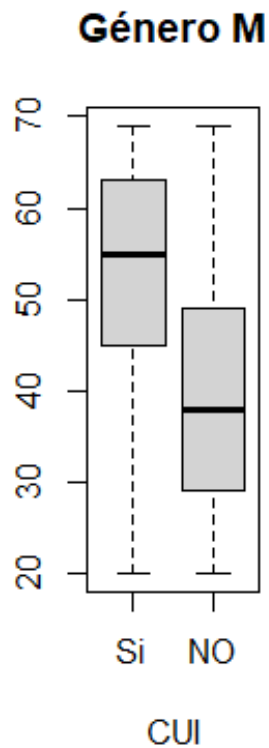


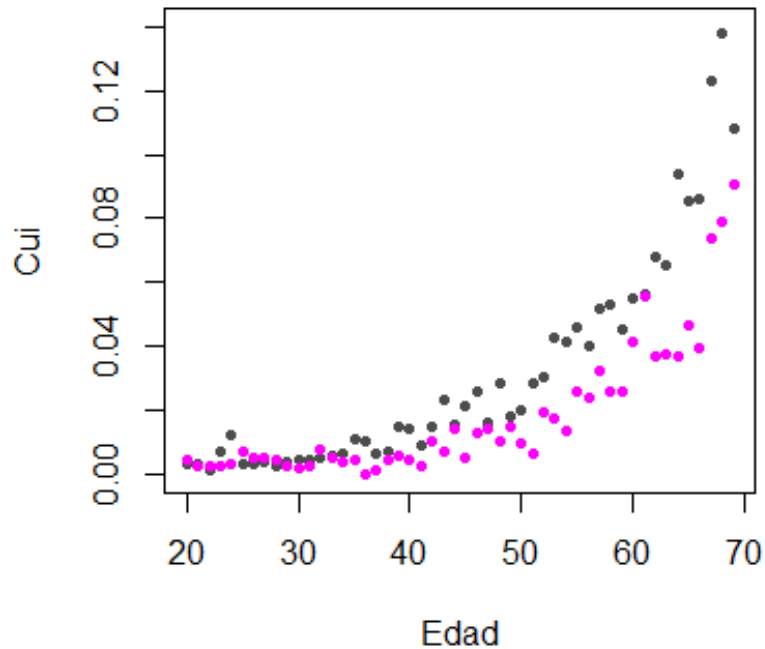
Gráficos



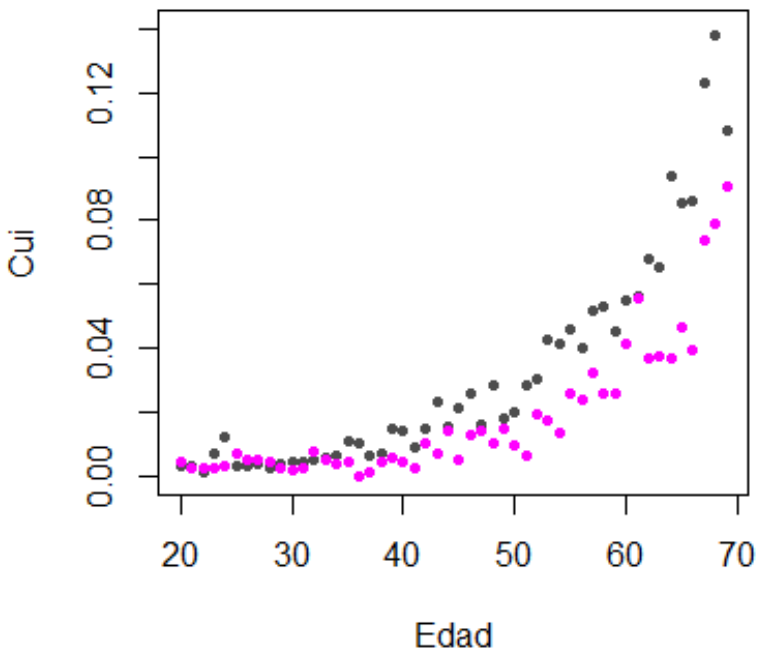
Gráficos

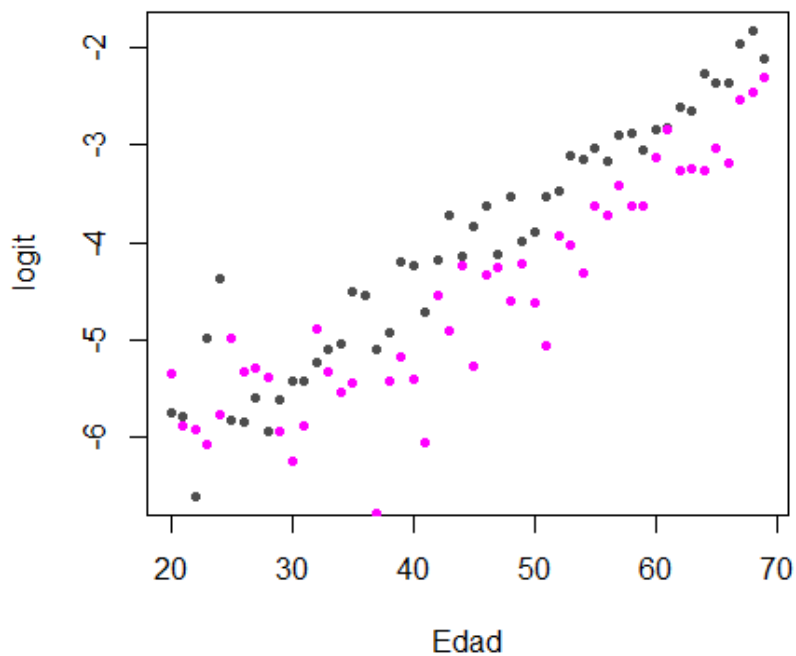

Género F
20
30
40
50
60
70
Si
NO
CUI


Género M
20
30
40
50
60
70
Si
NO
CUI


Cui
0.00
0.04
0.08
0.12
20
30
40
50
60
70
Edad

Gráficos


Cui
0.00
0.04
0.08
0.12
20
30
40
50
60
70
Edad


logit
-2
-3
-4
-5
-6
20
30
40
50
60
70
Edad

Ajustamos otro modelo usando Edad y Género

Llamemos

$$p(\text{Edad}) = \mathbb{P}(\text{CUI} = Si \mid (\text{Edad}, \text{Genero}))$$

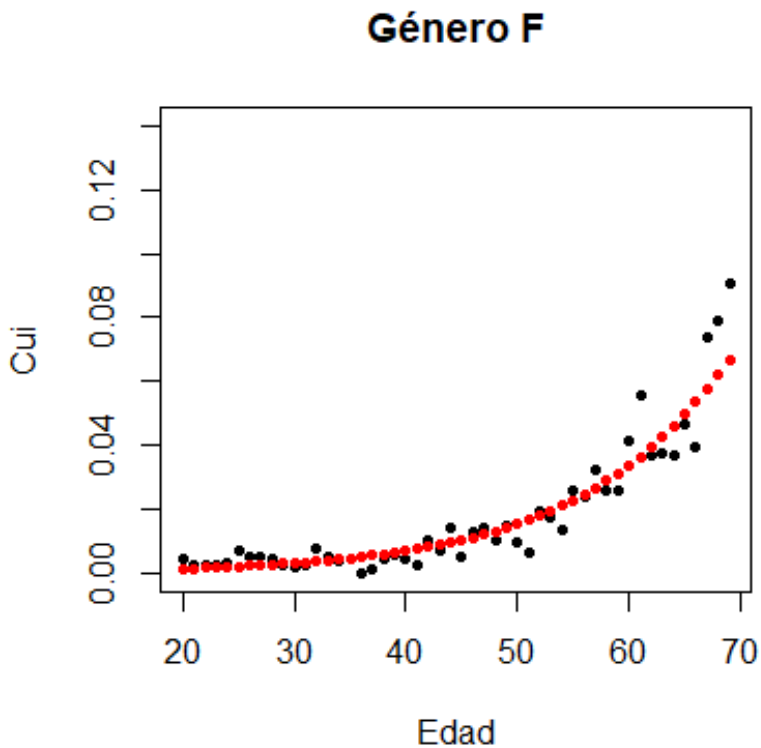
Ajustamos el modelo

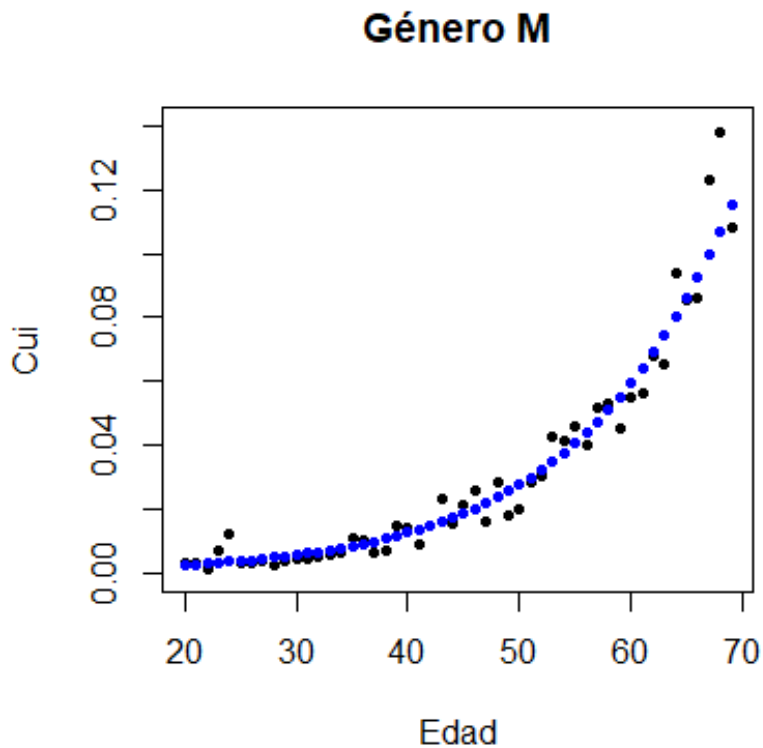
$$logit(p(\text{Edad}_i, \text{Genero}_i)) = \beta_1 + \beta_2\ \text{Edad}_i + \beta_3\ \text{Genero}_i$$

¿Qué efecto tiene el género en este modelo?

```
proba.hat=glm(cui~edad.dec+genero,family=binomial)$fitted

##########  promedio x genero
datosf<- filter(datos,genero==1)
datosm<- filter(datos,genero==0)

promediosf<- datosf%>% group_by(edad.dec) %>% summarize(mean(cui))
promediosm<- datosm%>% group_by(edad.dec) %>% summarize(mean(cui))
```

Gráficos


Género F
Cui
0.00
0.04
0.08
0.12
20
30
40
50
60
70
Edad


Género M
Cui
0.00
0.04
0.08
0.12
20
30
40
50
60
70
Edad