Modelos y métodos con variables dependientes

Máster Universitario en
Ingeniería Computacional y Sistemas Inteligentes
UPV/EHU

Métodos con variables dependientes - Indice

Métodos con variables dependientes - Indice

Introducción

Clasificación de las variables:

- según el valor que toman,
  cuantitativas, binarias, cualitativas.

- según lo que describen,
  1. *dependiente / respuesta,*
  2. *independiente / predictor,*
  3. *control / covariable / factor de confusión.*

  Esta clasificación depende más de los objetivos del estudio que de la naturaleza de las variables.

Introducción - Clasificación de las variables

Ej. 1

Y: Precio de la casa
X: Área de la casa (m^2)
C: Zona de la ciudad

Ej. 2

Y: Padecer una enfermedad
X: Ser fumador
C: sexo

$$\begin{array}{rcl} Y & \sim & X \\ \text{Var. DEPENDIENTE} & \sim & \text{Var. INDEPENDIENTE} \end{array}$$

$$\begin{array}{rcl} Y & \sim & X \\ \text{Var. DEPENDIENTE} & \sim & \text{Var. INDEPENDIENTE} \\ & \uparrow & \\ & C, & \text{Factor de CONFUSION} \end{array}$$

Modelos y métodos

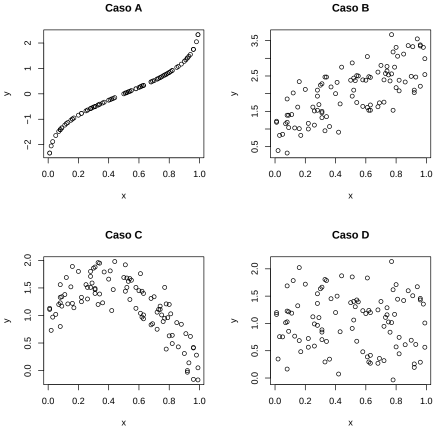
Consideramos la situación $Y \sim X$.

$X \backslash Y$	Cuantitativa	Cualitativa
Cuant.	Modelos regresión Árboles regresión	Análisis discriminante Clasificación supervisada
Cuali.	ANOVA	Chi- cuadrado

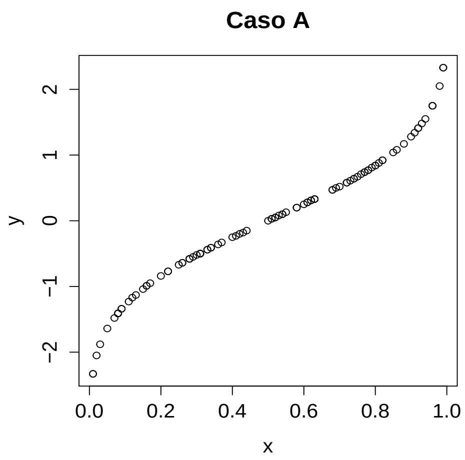
Métodos con variables dependientes - Indice

Introducción - Diferentes situaciones

Tenemos 2 variables cuantitativas X e Y.
Queremos ver cómo es Y dependiendo de X: $Y \sim X$

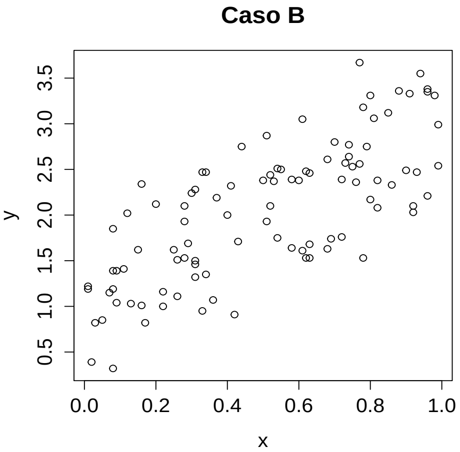


Introducción - Diferentes situaciones



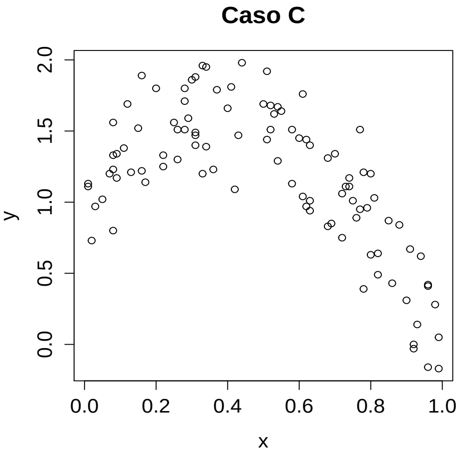
Volver

Introducción - Diferentes situaciones

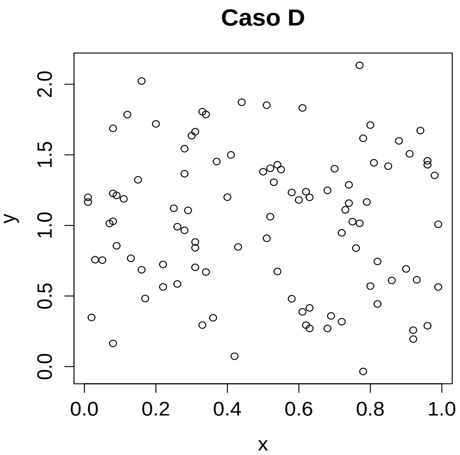


◂ Volver

Introducción - Diferentes situaciones

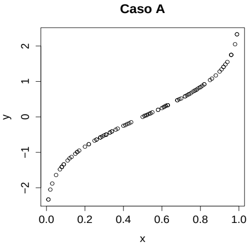


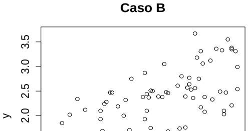
Introducción - Diferentes situaciones



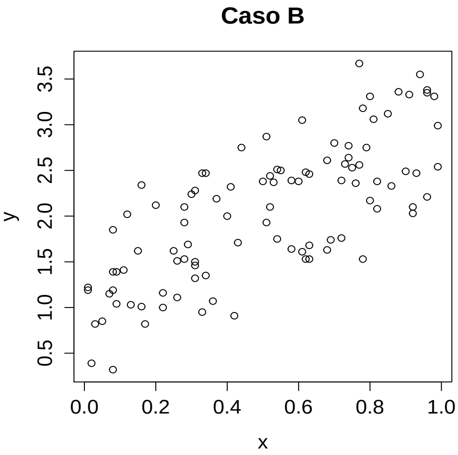
Introducción - Diferentes situaciones

¿ Cuál de las situaciones presentadas es adecuada para modelizar con un modelo de regresion lineal?





Introducción - Diferentes situaciones



Modelos de regresión

Sean X e Y dos variables cuantitativas. Los modelos de regresión de Y sobre X tienen en general el siguiente aspecto:

$$Y = \underset{\text{funcional}}{\underset{\text{comp.}}{f(X)}} + \underset{\text{aleatorio}}{\underset{\text{comp.}}{\epsilon}}$$

Y: var. dependiente ;
X: var. independiente ;
ϵ: error.

- ¿Cómo se relacionan las variables en el **caso A**? caso A
  ¿Te parece que está presente el componente aleatorio?
  $\Longrightarrow$Modelo determinista.
- ¿Cómo se relacionan las variables en el **caso B**? caso B
  $\Longrightarrow$Modelo de regresión.

Modelos de regresión - Ejemplos

Regresión simple (una única variable independiente)

- Contenido de hierro del suelo en función del porcentaje de materia orgánica del suelo.
- Distancia de frenado de un avión en función de la velocidad.
- Estudio del peso en función de la altura en personas adultas.

Regresión mútliple (más de una variable independiente)

- Volumen de venta de cierta bebida refrescante en función del tipo de envase y la temperatura.
- Estudio del peso en función de la altura y la edad en niños/as.

Modelos de regresión

Objetivos principales:

- Cuantificar la posible asociación entre las variables dependientes e independientes.
- Identificar de entre todas las variables independientes aquellas que están asociadas con la variable dependiente.
- Hacer predicciones.

Nota: Asociación no equivale a causalidad.

Métodos con variables dependientes - Indice

Modelo de regresión lineal simple

Supongamos que se han medido las variables **cuantitativas** X e Y sobre n individuos.
Es decir, tenemos n pares de valores $(x_1, y_1), \dots (x_n, y_n)$ donde:

- x_i es la respuesta del individuo i para la variable X,
- y_i es la respuesta del individuo i para la variable Y,
  $i = 1, \dots, n$.

Modelo de regresión lineal simple

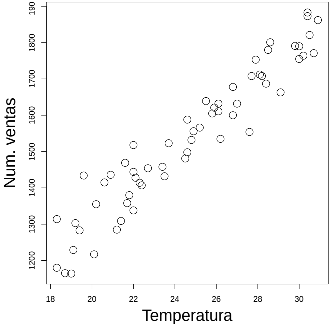
Por ejemplo,

un establecimiento quiere saber si el número de helados que vende al día está relacionado con la temperatura máxima (^oC) del día.

$$Y \equiv \text{“Número de ventas en un día”}$$

$$X \equiv \text{“Temperatura máxima del día”}$$

Día	Temp. x_i	Ventas y_i
1	26.1	1611
2	26.8	1600
3	28.4	1687
4	21.4	1309
⋮		



Modelo de regresión lineal simple

Regresión de Y sobre X

Los valores medios de Y están relacionados con X linealmente $\Rightarrow$
$E(Y|X = x) = \beta_0 + \beta_1 x$

Por tanto, escribimos el modelo como:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

Suposiciones básicas del modelo:

- $E(\epsilon_i) = 0, \quad i = 1, \ldots, n.$
- $VAR(\epsilon_i) = \sigma^2, \quad i = 1, \ldots, n.$
- $E(\epsilon_i \epsilon_j) = 0, \quad i \neq j.$
- Para hacer inferencia necesitamos además que $\epsilon_i \sim N(0, \sigma^2)$

Parámetros del modelo

- ¿ Cuántos parámetros tiene el modelo?

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \\
VAR(\epsilon_i) &= \sigma^2, \quad i = 1, \ldots, n
\end{aligned}
$$

- ¿ Qué información dan los parámetros?
- ¿Cómo se estiman los parámetros?

Interpretación del modelo de regresión lineal

Supongamos que ya hemos estimado los parámetros β_0, β_1 y σ^2 con b_0, b_1 y s^2, respectivamente.

Interpretación del modelo:

- Estimación del valor medio de Y supuesto que $X = x_i$,

$$\hat{y}_i = b_0 + b_1 x_i$$

- Estimación de la variabilidad de Y supuesto que $X = x_i$,

$$s^2$$

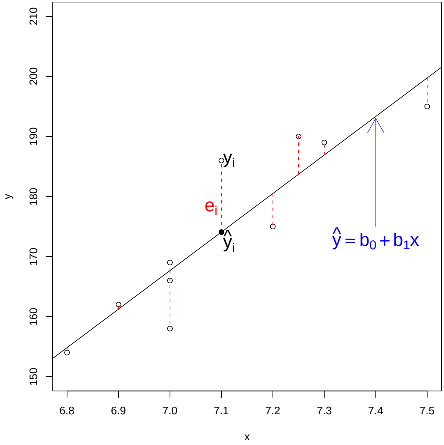
- Estimación del error, i.e., **residuo**,

$$e_i = y_i - \hat{y}_i$$

Estimación de los parámetros

Estimación por mínimos cuadrados

De entre todas las posibles rectas,
¿cuál es la que mejor se aproxima a los datos?



Minimizar

$$
\begin{aligned}
RSS(b_0, b_1) &= \sum_i e_i^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 \\
&= \sum_i (y_i - b_0 - b_1 x_i)^2
\end{aligned}
$$

Estimación de los parámetros

La estimación por mínimos cuadrados ofrece los siguientes estimadores:

- Pendiente de la recta de regresión:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{SXY}{SXX} \quad \text{(Notación)} \end{aligned}$$

- Término independiente de la recta de regresión:

$$b_0 = \bar{y} - b_1\bar{x}$$

Estimación de los parámetros

- Variabilidad del modelo o de los residuos:

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \\
&= \frac{RSS}{n-2} \quad \text{(Notación)}
\end{aligned}
$$

- **Definición** Se le llama error típico de regresión o error típico de estimación a

$$
s = \sqrt{\frac{RSS}{n-2}}
$$

Ejemplo - ventahelados.dat

Retomamos el ejemplo que relaciona la venta de helados con la temperatura.
Las estimaciones por mínimos cuadrados que nos da R son:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 339.405     48.527   6.994    3e-09 ***
temp         48.545      1.946  24.948   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 57.07 on 58 degrees of freedom
Multiple R-squared:  0.9148, Adjusted R-squared:  0.9133
F-statistic: 622.4 on 1 and 58 DF,  p-value: < 2.2e-16
```

Modelo calculado:

$$\widehat{ventas} = 339.405 + 48.545\,temp\,, \quad s = 57.07$$

Ejemplo - ventahelados.dat

Modelo calculado:

$\widehat{ventas} = 339.405 + 48.545 temp\,, \quad s = 57.07$

Según este modelo:

1. Para un día con temperatura máxima de $25^o C$ se espera una venta media de 1553 helados
   $(1553.03 = 339.405 + 48.545 \times 25)$
2. Con el incremento de $1^o C$ en la temperatura, se estima que incrementa en $b_1 = 48.545$ unidades la venta media.
3. ¿Cómo se interpretaría el valor $b_0 = 339.405$? ¿Tiene sentido?

Recta de regresión. Algunas propiedades

Prop. 1 Consideramos el centro de gravedad $(\bar{x}, \bar{y})$.
La recta de regresión pasa por el centro de gravedad.

Prop. 2 Ya sabemos que el coeficiente de correlación entre X y Y mide la asociación lineal entre las dos variables. Se tiene que

$$b_1 = r\frac{\sqrt{SYY}}{\sqrt{SXX}},$$

donde r es el coeficiente de correlación entre X y Y.

Descomposición de la variabilidad

Variabilidad de Y:

$$SYY = \sum_i (y_i - \bar{y})^2$$

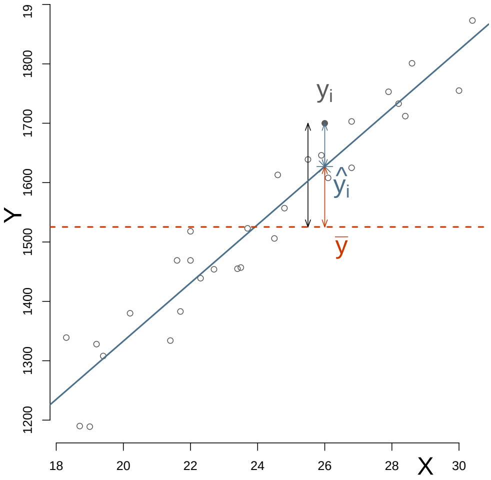
Descomposición de la variabilidad de la var. dependiente

$$\begin{array}{ccccc} \sum_i (y_i - \bar{y})^2 & = & \sum_i (\hat{y}_i - \bar{y})^2 & + & \sum_i (y_i - \hat{y}_i)^2 \\ \\ SYY & = & SSreg & + & RSS \\ \text{Var. TOTAL} & & \text{Var. EXPLICADA} & & \text{Var. RESIDUAL} \end{array}$$

(Notación)

Descomposición de la variabilidad

Miremos la descripción gráfica de $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$



Coeficiente de determinación

Definición

Se define el coeficiente de determinación del modelo como:

$$R^2 = \frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY}$$

- R^2 se interpreta como el % de variabilidad de Y explicado a través del modelo.
- $\sqrt{R^2} = |r|$

Tabla ANOVA - Test de regresión

Fuentes de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F
Regresión	1	$SSreg$	$SSreg/1$ $= MSreg$	$MSreg/RMS$
Residual	$n-2$	RSS	$RSS/(n-2)$ $= RMS$	
Total	$n-1$	SYY		

Ejemplo - ventahelados.dat

```
 helados <- read.table("ventahelados.dat", header=TRUE)
 rl <- lm(ventas ~ temp, data=helados)
 summary(rl)

Call:
lm(formula = ventas ~ temp, data = helados)

Residuals:
     Min       1Q   Median       3Q      Max
-125.240  -40.490    2.547   35.396  143.118

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  339.405     48.527   6.994    3e-09 ***
temp          48.545      1.946  24.948   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 57.07 on 58 degrees of freedom
Multiple R-squared:  0.9148, Adjusted R-squared:  0.9133
F-statistic: 622.4 on 1 and 58 DF,  p-value: < 2.2e-16
```

Ejemplo - ventahelados.dat

```
rl <- lm(ventas ~ temp, data=helados)
anova(rl)
```

```
Analysis of Variance Table

Response: ventas
          Df  Sum Sq Mean Sq F value    Pr(>F)
temp       1 2027117 2027117  622.42 < 2.2e-16 ***
Residuals 58  188895    3257
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- ¿Cuánto vale SYY?
- ¿Cuánto vale R^2?

Métodos con variables dependientes - Indice

Modelo de regresión lineal múltiple

Supongamos que se han medido las variables **cuantitativas** $X_1, \ldots, X_p$ e Y sobre n individuos.

Tenemos n $(p+1)$-tuplas de valores $(x_{11}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, \ldots, x_{np}, y_n)$ donde:

- x_{ij} es la respuesta del individuo i para la variable X_j, $j = 1, \ldots p$,
- y_i es la respuesta del individuo i para la variable Y, $i = 1, \ldots, n$.

Modelo de regresión lineal múltiple

Regresión de Y sobre $X_1, \ldots, X_p$

Los valores medios de Y están relacionados linealmente con $X_1, \ldots, X_p \;\Rightarrow$
$E(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Por tanto, escribimos el modelo como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$

Suposiciones básicas del modelo:

- $E(\epsilon_i) = 0, \quad i = 1, \ldots, n.$
- $VAR(\epsilon_i) = \sigma^2, \quad i = 1, \ldots, n.$
- $E(\epsilon_i \epsilon_j) = 0, \quad i \neq j.$
- Para hacer inferencia necesitamos además que $\epsilon_i \sim N(0, \sigma^2)$

Parámetros del modelo

- Los parámetros del modelo son:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$
$$VAR(\epsilon_i) = \sigma^2, \quad i = 1, \ldots, n$$

Interpretación del modelo de regresión múltiple

Supongamos que ya hemos estimado los parámetros β_0, $\beta_1, \ldots, \beta_p$ y σ^2 con b_0, $b_1, \ldots, b_p$ y s^2, respectivamente.

Interpretación del modelo:

- estimación del valor medio de Y supuesto que $X_1 = x_{i1}, \ldots, X_p = x_{ip}$,

$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$$

- estimación de la variabilidad de Y supuesto que $X_1 = x_{i1}, \ldots, X_p = x_{ip}$,

$$s^2$$

- residuos,

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

Estimación de los parámetros

Estimación por mínimos cuadrados

Hallar $b_1, \ldots, b_p$ y b_0 de manera que se minimice:

$$\begin{aligned} RSS(b_0, b_1, \ldots, b_p) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))^2 \end{aligned}$$

$$\Downarrow$$

Ecuaciones NORMALES

Estimación de los parámetros

Matricialmente, definimos:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}.$$

$$\text{Luego,} \qquad \hat{Y} = \mathbf{X}\mathbf{b}.$$

Objetivo: Minimizar $RSS(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$

Estimadores mínimo cuadráticos:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$$

Estimación de los parámetros

Podemos poner los residuos como:

$$\begin{aligned} \mathbf{e} &= (e_1, \ldots, e_n)' \\ &= \mathbf{Y} - \hat{\mathbf{Y}} \end{aligned}$$

Estimaremos variabilidad del modelo con los residuos:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(p+1)} \\ &= \frac{\mathbf{e}'\mathbf{e}}{n-(p+1)} \\ &= \frac{RSS}{n-p-1} \end{aligned}$$

Definición Se le llama error típico de regresión a $s = \sqrt{\frac{RSS}{n-p-1}}$.

Matriz “hat”, H

Podemos escribir las estimaciones como:

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{Xb} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{HY}\end{aligned}$$

definiendo $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Definición Se llama *matriz hat* a $\mathbf{H}$.

Nota: Esta matriz está muy presente en la *Evaluación del modelo de regresión*.

Descomposición de la variabilidad

Variabilidad de Y:

$$SYY = \sum_i (y_i - \bar{y})^2$$

Descomposición de la variabilidad de la var. dependiente

$$\begin{array}{ccccc} \sum_i (y_i - \bar{y})^2 & = & \sum_i (\hat{y}_i - \bar{y})^2 & + & \sum_i (y_i - \hat{y}_i)^2 \\ & & & & \\ SYY & = & SSreg & + & RSS \\ \text{Var. TOTAL} & & \text{Var. EXPLICADA} & & \text{Var. RESIDUAL} \end{array}$$

Se recupera la misma
descomposición de la variabilidad
que en regresión simple!

Tabla ANOVA - Test de regresión

Fuentes de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F
Regresión	p	$SSreg$	$SSreg/p$ $= MSreg$	$MSreg/RMS$
Residual	$n-(p+1)$	RSS	$RSS/(n-p-1)$ $= RMS$	
Total	$n-1$	SYY		

Coeficiente de determinación

Definición

Se define el coeficiente de determinación del modelo como:

$$R^2 = \frac{SSreg}{SYY}$$

- R^2 se interpreta como el % de variabilidad de Y explicado a través del modelo.

Definición

Se dice coeficiente de correlación múltiple a $\sqrt{R^2}$.

Coeficiente de determinación corregido

Definición

Se define el coeficiente de determinación corregido del modelo como:

$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

- $\bar{R}^2 \leq R^2$.

Ejemplo

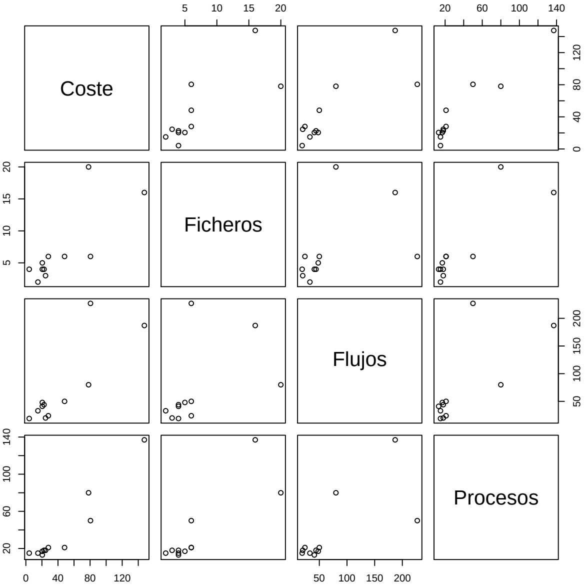
Los elementos de estructura básicos de un sistema de procesamiento de datos son tres:

- **Ficheros**. Forman un conjunto de registros permanentes en el sistema.
- **Flujos**. Interfaces de datos entre el sistema y el entorno.
- **Procesos**. Manipulaciones lógicas funcionales definidas sobre los datos.

Se estudian estos tres elementos a fin de estudiar el coste del desarrollo del software. Se han recogido los siguientes datos:

Coste	22.6	15.0	78.1	28.0	80.5	24.5	20.5	147.6	4.2	48.2	20.5
Ficheros	4	2	20	6	6	3	4	16	4	6	5
Flujos	44	33	80	24	227	20	41	187	19	50	48
Procesos	18	15	80	21	50	18	13	137	15	21	17

Ejemplo - continuación



Ejemplo - continuación

```
regCoste <- lm(Coste~Ficheros+Flujos+Procesos)
summary(regCoste)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.96178    5.60831   0.350  0.73678
Ficheros     0.11776    1.17665   0.100  0.92309
Flujos       0.17673   0.07144   2.474  0.04260 *
Procesos     0.79645  0.22042   3.613  0.00859 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error:  9.918 on 7 degrees of freedom
Multiple R-squared: 0.9615, Adjusted R-squared: 0.9449
F-statistic: 58.21 on 3 and 7 DF,  p-value: 2.578e-05
```

Ejemplo - continuación

El modelo calculado es:

$$\begin{aligned}
\widehat{coste} &= 1.96 + 0.12 \times \textit{Ficheros} + 0.18 \times \textit{Flujos} + 0.80 \times \textit{Procesos} \\
s &= 9.92 \\
R^2 &= 0.9615
\end{aligned}$$

Según este modelo:

1. El modelo explica el 96.15% de la variabilidad del coste del desarrollo del software.
2. Para procesamientos en los que intervienen 10 ficheros, 100 flujos y 75 procesos, se espera un coste medio de 80.80. $(80.80 = 1.96 + 0.11776 \times 10 + 0.17673 \times 100 + 0.79645 \times 75)$

Métodos con variables dependientes - Indice

Modelo Lineal General - motivación

Ya hemos visto el ejemplo en el que un establecimiento quiere saber si el número de helados que vende por día está relacionado con la temperatura máxima (^oC) del día.
Si además queremos ver si la venta depende de si el día es laboral, laboral víspera de festivo o festivo, tenemos las siguientes variables:

$$
\begin{aligned}
Y &\equiv \text{“Número de ventas en un día”} \\
X_1 &\equiv \text{“Temperatura máxima del día”} \\
X_2 &\equiv \text{“Tipo de día (1-laboral; 2-víspera festivo; 3-festivo)”}
\end{aligned}
$$

¿Te parece adecuado plantear el modelo
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$?

Modelo Lineal General

Construcción de las variables indicadoras

Como la variable cualitativa X_2 tiene 3 modalidades, generamos 2 variables indicadoras Z_1 y Z_2 de la siguiente manera:

$$Z_1 = \begin{cases} 1\,, \text{ si } X_2 = \text{ laboral} \\ 0\,, \text{en caso contrario} \end{cases} \qquad Z_2 = \begin{cases} 1\,, \text{ si } X_2 = \text{ vispera festivo} \\ 0\,, \text{en caso contrario} \end{cases}$$

Por ejemplo:

$$\begin{array}{c} X_2 \\ \hline 1 \\ 2 \\ 3 \\ 1 \\ 1 \\ \ldots \end{array} \quad \Leftrightarrow \quad \begin{array}{cc} Z_1 & Z_2 \\ \hline 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ \ldots & \end{array}$$

Modelo Lineal General

Por tanto, calcularemos el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \overbrace{\beta_2 Z_1 + \beta_3 Z_2} + \epsilon$$

Supongamos que hemos obtenido el siguiente ajuste:

$$\text{A1: } \hat{Y} = 400 + 48X_1 - 54Z_1 - 45Z_2$$

Entonces, hemos ajustado 3 rectas:

$$\text{A1: } \begin{cases} \hat{Y} = 400 - 54 + 48X_1\,, & \text{si laboral} \\ \hat{Y} = 400 - 45 + 48X_1\,, & \text{si vispera} \\ \hat{Y} = 400 + 48X_1\,, & \text{si festivo} \end{cases}$$

Modelo Lineal General

Algunos comentarios:

- La interpretación del modelo lineal general es la misma que el modelo de regresión múltiple.
- En el software R utilizaremos la funcion `glm()` y no hace falta generar las variables indicadoras.

Ejemplo - helados.dat

```
ml <- glm(ventas ~ temp + tipodia, data=helados)
summary(ml)

Call:
glm(formula = ventas ~ temp + tipodia, data = helados)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-107.183   -34.844     8.068    40.185   100.406

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  345.893     45.138   7.663 2.78e-10 ***
temp          47.655      1.827  26.086  < 2e-16 ***
tipodia2       8.443     20.735   0.407  0.68543
tipodia3      53.654     16.097   3.333  0.00153 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for gaussian family taken to be 2807.805)
```

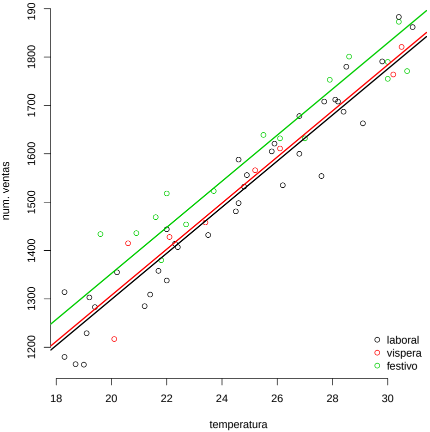
Ejemplo - helados.dat

Modelo ajustado:

$$\hat{y} = 345.9 \ + 47.7 \times temp\,, \quad \text{si laboral}$$

$$\hat{y} = 345.9 + 8.4 \ + 47.7 \times temp$$
$$= 354.3 \ + 47.7 \times temp\,, \quad \text{si vispera}$$

$$\hat{y} = 345.9 + 53.7 \ + 47.7 \times temp$$
$$= 399.5 \ + 47.7 \times temp\,, \quad \text{si festivo}$$



Métodos con variables dependientes - Indice

Evaluación de un modelo de regresión

Debemos comprobar las hipótesis básicas del modelo que hemos construido.

- **Linealidad**: ¿Parece razonble considerar que la media condicionada de Y es lineal respecto de X?
- **Variabilidad constante**: ¿Parece razonable asumir que los errores tienen una variabilidad constante?
- **Independencia de los errores**: ¿Parece razonable considerar que los errores son independientes entre sí?
- **Individuos influyentes - valores extremos**: ¿Hay algún individuo que modifique sustancialmente las estimaciones de los parámetros?

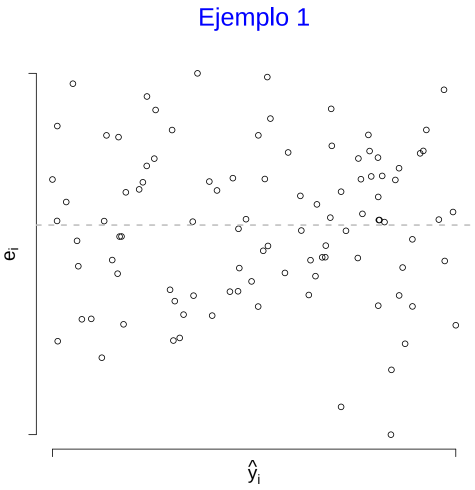
Estudio de los residuos

El estudio de los residuos nos ayudará a determinar si el modelo en cuestión es válido.

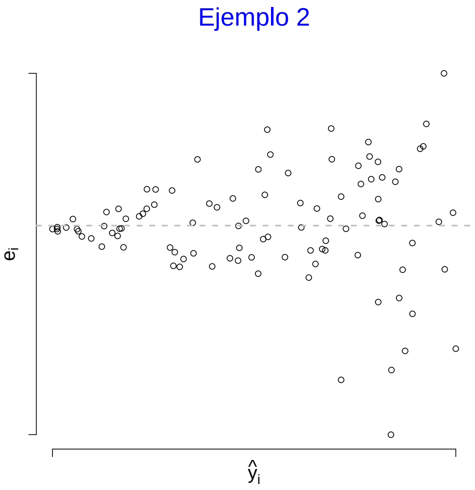
- Residuos: $e_i = y_i - \hat{y}_i \,, \quad i = 1, \ldots, n.$
- Residuos estandarizados: $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \,, \quad i = 1, \ldots, n$, donde h_{ii} son los elementos de la diagonal de la matriz H.

Si el modelo considerado es válido, los residuos no deben de mostrar ninguna pauta. En el caso de que los residuos muestren alguna pauta indica que el modelo ajustado no es válido y por tanto hay que replantearlo.

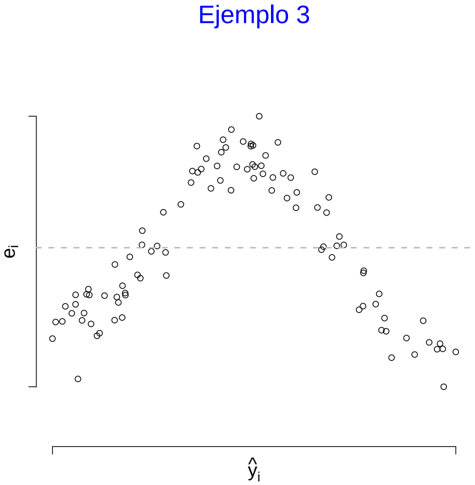
Diagnóstico del modelo
Representación gráfica de los residuos



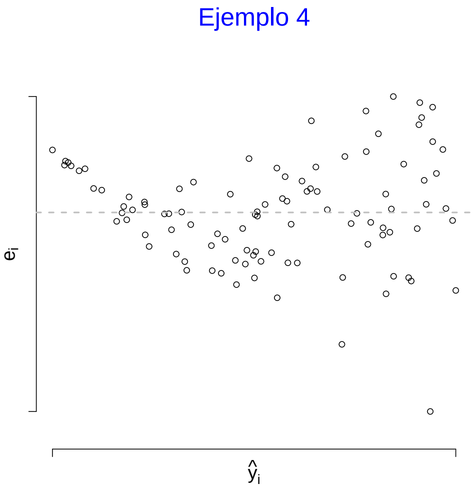
Diagnóstico del modelo
Representación gráfica de los residuos



Diagnóstico del modelo
Representación gráfica de los residuos



Diagnóstico del modelo
Representación gráfica de los residuos



Individuos influyentes

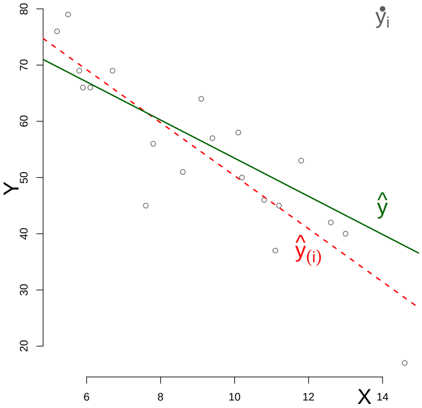
- Ajuste con todos los ind.

$$\hat{y} = b_0 + b_1 x$$

- Ajuste sin el individuo y_i

$$\hat{y}_{(i)} = b_{0(i)} + b_{1(i)} x$$

Por ejemplo,
si b_1 y $b_{1(i)}$ muy diferentes,
y_i es influyente.



Individuos influyentes - Distancia de Cook

Sean:

- $\mathbf{b} = (b_0, b_1, \ldots, b_p)'$ el vector con las estimaciones de los coeficientes
- $\mathbf{b}_{(i)} = (b_{0(i)}, b_{1(i)}, \ldots, b_{p(i)})'$ el vector con las estimaciones de los coeficientes una vez eliminado en individuo i

Se define la distancia de Cook para el individuo i como:

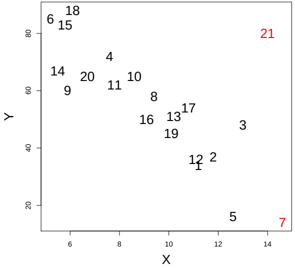
$$
\begin{aligned}
D_i &= \frac{(\mathbf{b}_{(i)} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_{(i)} - \mathbf{b})}{(p+1)s^2} \\
&= \frac{r_i^2}{k+1} \times \frac{h_{ii}}{1-h_{ii}}
\end{aligned}
$$

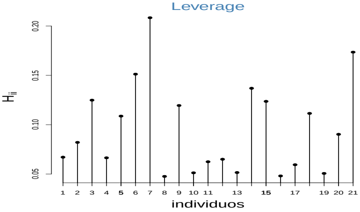
Valores altos de D_i pueden indicar individuos influyentes.

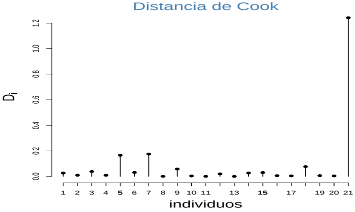
Individuos influyentes

```
m <- lm(y ~ x)
hatvalues(m)
cooks.distance(m)
```

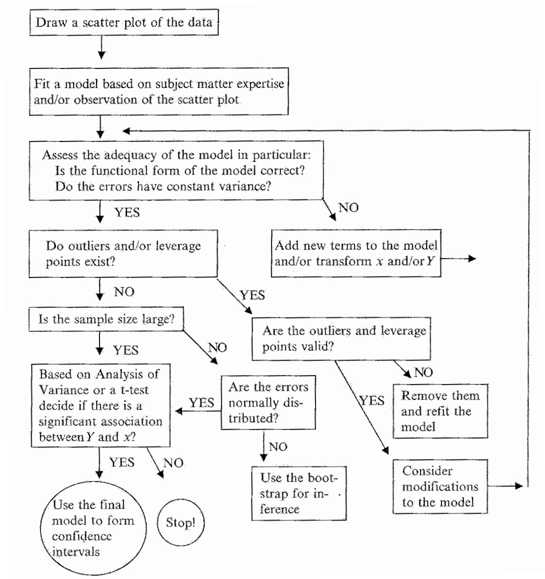
	h_{ii}	D_i
1	0.07	0.03
2	0.08	0.01
...		
7	0.21	0.18
8	0.05	0.00
...		
21	0.17	1.24







Esquema general



S.J.Sheather (2009) *A Modern Approach to Regression with R*. Springer (pág 103)

Métodos con variables dependientes - Indice

Métodos con variables dependientes - Indice

Análisis Discriminante (Clasificación supervisada)

- Dada una imagen digital de los caracteres B ó 8, identificarlo correctamente .
- Conocidas ciertas caracteristicas y síntomas de la enfermedad de un paciente, idenficar el correcto diagnóstico.

Formalización del problema

Supongamos que tenemos Ω clasificado en g grupos excluyentes y exhaustivos, $G_1, \ldots, G_g$.
Consideraremos Y, de manera que $Y(i) = l$ si y sólo si $i \in G_l$.

Por otra parte, tenemos $X_1, \ldots, X_p$ variables explicativas.

Objetivo pricipal:

- Dados valores de $\mathbf{x} = (x_1, \ldots, x_p)'$, predecir el grupo de pertenencia.

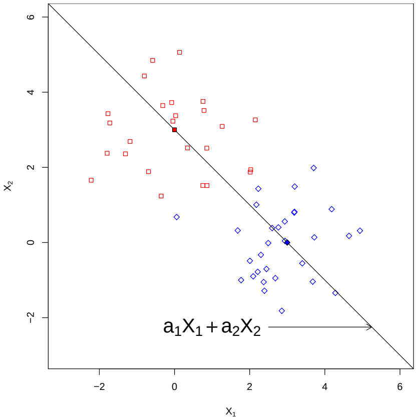
Métodos con variables dependientes - Indice

Discriminador lineal de Fisher ($g = 2$)

Supongamos el caso $g = 2$ con $X_1, \ldots, X_p$ variables cuantitativas. Necesitamos una regla que nos permita clasificar un objeto $\mathbf{x}$ en G_1 ó G_2.

Idea intuitiva para buscar la regla

Buscar la recta $a_1 X_1 + a_2 X_2$ que separa al máximo las medias de cada grupo.



Discriminador lineal de Fisher ($g = 2$)

Sean n_l, $\bar{\mathbf{x}}_l$ y $\mathbf{S}_l$, número de objetos, media y matriz de varianzas-covarianzas del grupo G_l, respectivamente ($l = 1, 2$).

Consideramos los coeficientes $\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ donde $\mathbf{S} = ((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2)/(n - 2)$. Dado un objeto $\mathbf{x}$, calculamos su proyección en la recta: $t = \mathbf{a}'\mathbf{x}$.

Regla de clasificación:

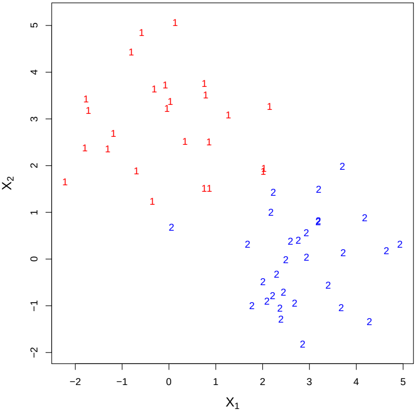
$$\text{Clasificamos } \mathbf{x} \text{ en } G_1 \quad \text{sii} \quad |t - \bar{t}_1| < |t - \bar{t}_2|.$$

Ejemplo

Se han medido dos variables ficticias en dos grupos ($n_1 = 25$, $n_2 = 30$).

Aspecto del conjunto de datos:

```
      X1    X2  group
[1,] 0.753 1.518 1
[2,] 2.014 1.873 1
[3,] 3.198 1.490 2
[4,] 2.028 1.937 1
[5,] 2.938 0.043 2
      .
      .
      .
```



Debemos calcular los coeficientes **a**.

Ejemplo (cont.)

- Características de los dos grupos:
  Medias

```
>        m1
[1,] -0.0343801
[2,]  2.9047747
```

```
>         m2
[1,]  2.83761061
[2,] -0.02712366
```

Matrices de varianza-covarianza

```
>              S1
[1,]  1.5323224 -0.1002049
[2,] -0.1002049  1.0861350
```

```
>              S2
[1,] 0.97220069 0.08280868
[2,] 0.08280868 0.89772440
```

Matriz de varianzas-covarianzas común para los dos grupos

```
>                S
[1,]  1.225841     -6.537e-05
[2,] -6.537e-05    0.9830424
```

- Los coeficientes: `a = [ -2.343, 2.982]`

Ejemplo (cont.)

Buscamos las proyecciones de cada individuo.

Para el individuo 1: $\mathbf{x}_1 = (0.753, 1.518)'$
$t_1 = -2.343 \cdot 0.753 + 2.972 \cdot 1.518 = 2.763$

De manera similar para el resto:

```
       X1      X2   group      t
[1,]  0.753  1.518  1       2.762
[2,]  2.014  1.873  1       0.868
[3,]  3.198  1.490  2      -3.048
[4,]  2.028  1.937  1       1.026
[5,]  2.938  0.043  2      -6.754
       ...
```

Por tanto, $\bar{t}_1 = 8.744$, y $\bar{t}_2 = -6.729$

Según esta regla,
¿en qué grupo se clasifica un individuo con características
$\mathbf{x} = (1, 0)'$?

Diferentes puntos de vista para abordar el problema

Sea **x** el objeto a clasificar en alguna de las clases G_1, G_2
Medir $d(x, G_i)\,, \quad i = 1, 2$

Criterio geométrico

Consiste en asignar **x** a la clase más cercana,

- Asignar **x** a G_1 si $d(x, G_1) < d(x, G_2)$
- Asignar **x** a G_2 en caso contrario

Pero...

¿Cómo se puede medir $d(x, G_i)\,,\ i = 1, 2$?

Diferentes puntos de vista para abordar el problema

Supongamos que las variables $X_1, \ldots, X_p$ tienen una distribución de probabilidad según la función de densidad $f(x_1, \ldots, x_p; \theta)$.
En G_1, $f(x_1, \ldots, x_p; \theta_1)$
En G_2, $f(x_1, \ldots, x_p; \theta_2)$

Sea $\mathbf{x} = (x_1, \ldots, x_p)'$ el objeto a clasificar en G_1, o G_2.
La probabilidad o verosimilitud de la observación $\mathbf{x}$ en la clase G_i es

$$L_i(\mathbf{x}) = f(x_1, \ldots, x_p; \theta_i) \quad (i = 1, 2).$$

Regla de máxima verosimilitud

Consiste en asignar $\mathbf{x} = (x_1, \ldots, x_p)'$ a la clase tal que la verosimilitud de $(x_1, \ldots, x_p)$ es mayor, es decir,

- Asignar $\mathbf{x}$ a G_i si $L_i(\mathbf{x}) = \max\{L_1(\mathbf{x}), L_2(\mathbf{x})\}$

Diferentes puntos de vista para abordar el problema

Supongamos que se conoce la probabilidad $q_i = P(G_i)$ de que un objeto pertenezca a G_i.
Dado $\mathbf{x} = (x_1, \ldots, x_p)'$, de verosimilitud $L_i(\mathbf{x}) = f(x_1, \ldots, x_p; \theta_i)$, la probabilidad, *a posteriori*, viene dada por la regla de Bayes,

$$P(G_i|\mathbf{x}) = \frac{q_i L_i(\mathbf{x})}{q_1 L_1(\mathbf{x}) + q_2 L_2(\mathbf{x})} \quad (i = 1, 2)$$

Regla de Bayes

Consiste en asignar $\mathbf{x} = (x_1, \ldots, x_p)'$ a la clase tal que la probabilidad a posteriori es mayor,

- Asignar $\mathbf{x}$ a G_i si $q_i L_i(\mathbf{x}) = \max\{q_1 L_1(\mathbf{x}), q_2 L_2(\mathbf{x})\}$

Discriminador lineal de Fisher ($g = 2$)

Modos equivalentes de expresar el discriminador lineal de Fisher.

- Se define $LDF(\mathbf{x}) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\left[\mathbf{x} - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2\right]$, entonces,

$$\text{clasificamos } \mathbf{x} \text{ en } G_1 \quad \text{sii} \quad LDF(\mathbf{x}) > 0.$$

- Se considera la distancia Mahalanobis (al cuadrado) al centro de cada grupo:

$$D_l^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_l)'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_l) \quad l = 1, 2,$$

  entonces,

$$\text{clasificamos } \mathbf{x} \text{ en } G_1 \quad \text{sii} \quad D_2^2(\mathbf{x}) > D_1^2(\mathbf{x}).$$

Métodos con variables dependientes - Indice

Evaluación de una regla discriminante

Una estimación de la tasa de estimación errónea provee de una medida cuantitativa del poder de discriminación de una regla discriminante.

- **Error aparente** o **Estimación por resubstitución**: Consiste en estimar la función discriminante con todos los objetos, usar esta regla para clasificar los objetos, y finalmente calcular la tasa de estimación errónea. **Ojo!**
- **Método de la escisión**: Consiste en dividir la muestra en dos submuestras de manera que con una submuestra se estima la función discriminante para clasificar los objetos de la otra submuestra y así calcular la tasa de estimación errónea.
- **Validación Cruzada**: Consiste en estimar la función discriminante dejando fuera un objeto, y luego usar la regla para clasificar el objeto. Se repite el procedimiento para todas los objetos y se calcula la tasa de clasificación errónea.